



(11) **EP 4 424 698 A2**

(12) **EUROPEAN PATENT APPLICATION**

(43) Date of publication:  
**04.09.2024 Bulletin 2024/36**

(51) International Patent Classification (IPC):  
**C07K 1/113 (2006.01)**

(21) Application number: **24185760.6**

(52) Cooperative Patent Classification (CPC):  
**C07K 1/1077; A61P 31/12; C07K 1/113**

(22) Date of filing: **29.07.2022**

(84) Designated Contracting States:  
**AL AT BE BG CH CY CZ DE DK EE ES FI FR GB  
GR HR HU IE IS IT LI LT LU LV MC MK MT NL NO  
PL PT RO RS SE SI SK SM TR**

(30) Priority: **30.07.2021 US 202163227820 P**  
**29.10.2021 US 202163273858 P**

(62) Document number(s) of the earlier application(s) in  
accordance with Art. 76 EPC:  
**22751458.5 / 4 175 968**

(71) Applicant: **Helaina, Inc.**  
**New York, NY 10010 (US)**

(72) Inventors:  
• **BOTERO BESADA-LOMBANA, Pamela**  
**New York, 10010 (US)**

• **KATZ, Laura**  
**New York, 10010 (US)**

(74) Representative: **Bohmann, Armin K.**  
**Bohmann**  
**Anwaltssozietät**  
**Nymphenburger Straße 1**  
**80335 München (DE)**

Remarks:

- The complete document including Reference Table(s) and the Sequence Listing(s) can be downloaded from the EPO website
- This application was filed on 01-07-2024 as a divisional application to the application mentioned under INID code 62.

(54) **METHODS AND COMPOSITIONS FOR PROTEIN SYNTHESIS AND SECRETION**

(57) Disclosed herein, in some aspects, are synthetic secretion signal peptides. Also disclosed are nucleic acid molecules encoding such signal peptides, in some cases operably linked to a protein coding sequence, as well as cells comprising such nucleic acid molecules. Further disclosed are methods for secreting a polypeptide com-

prising expressing in a cell a signal peptide of the disclosure linked to the polypeptide. Certain aspects include proteins (e.g., human milk proteins) produced by such methods, as well as compositions comprising such proteins.

**EP 4 424 698 A2**

**Description****CROSS REFERENCE TO RELATED APPLICATIONS**

**[0001]** This application claims priority to and the benefit of U.S. Provisional Application No. 63/227,820, filed July 30, 2021, and U.S. Provisional Application No. 63/273,858, filed October 21, 2021, which are hereby incorporated by reference in their entirety.

**SEQUENCE LISTING**

**[0002]** The instant application contains a Sequence Listing which has been submitted in XML format and is hereby incorporated by reference in its entirety. Said XML copy, created on July 76, 2022, is named HELA\_P0005WO\_Sequence\_Listing.xml and is 61,471 bytes in size.

**BACKGROUND****I. Technical Field**

**[0003]** Aspects of this invention relate to at least the fields of microbiology, genetics, and biotechnology.

**II. Background**

**[0004]** Yeast is a desirable host for production of recombinant proteins due to its rapid growth, its ability to reach high cell densities, to grow on defined minimal media, achieve high protein yields and conduct eukaryotic post-translational modifications. The most relevant yeast for protein production is *Pichia pastoris* (*Komagataella pastoris*, *Komagataella phaffii*) due to the wide availability of genomic information and molecular tools for genomic manipulation. These have enabled the use of *Pichia pastoris* for production of GRAS ingredients based on the FDA criteria.

**[0005]** For diverse biotechnological applications it is often preferred to produce proteins that are secreted to the growth medium to ease recovery. *Pichia pastoris* is capable of secreting active recombinant proteins, while maintaining low-level secretion of endogenous proteins.

**[0006]** In eukaryotes, secreted proteins are first targeted from the cytoplasm to the lumen endoplasmic reticulum (ER) via translocation. Translocation into the ER can take place either post-translationally (i.e., once the polypeptide chain has been synthesized) or co-translationally (i.e., during mRNA translation into its amino acid sequence). Post-translational translocation requires chaperones that maintain the polypeptide chain in a loose conformation in the cytosol as well as the action of the ER-resident chaperone Kar2, which acts as a molecular ratchet. Consequently, this process can be hindered by partially folded domains and/or cytosolic aggregation. Therefore, for biotechnological applications it is desirable to promote co-translational translocation. Once in the ER, proteins are glycosylated, their disulfide bonds are isomerized, and they fold to their native state. Proteins that are successfully folded then transit to the Golgi complex, where further glycosylation takes place before being packed into secretory granules that fuse to the cell membrane, releasing the protein to the extracellular milieu.

**[0007]** Targeting of the proteins to the secretory pathway is mediated by secretion peptides. The most widely used in *Pichia pastoris* is the leader peptide of the mating factor alpha of *S. cerevisiae*. It is comprised of two distinct regions: ii) the first 19 amino acid pre-region that promotes post-translational translocation and is cleaved upon ER entry ii) a 70 amino acid pro-segment that serves as an ER-to-Golgi export signal and it is cleaved in the Golgi Apparatus at the dibasic amino acid cleavage site KR.

**[0008]** There exists a need for synthetic secretion signal peptides leading to higher extracellular production of proteins.

**SUMMARY**

**[0009]** Aspects of the present disclosure address certain needs by providing novel secretion signal peptides effective in improving extracellular production of proteins, including mammalian proteins such as, for example, human milk proteins. Certain aspects of the disclosure are based, at least in part, on the development of signal peptides generated from the in-frame fusion of 1) pre-secretion peptides of *P. pastoris* from either i) the alpha subunit of the oligosaccharyltransferase complex of the ER lumen (Ost1) or ii) the GPI-anchored protein Pst1 with 2) the pro-region of either i) the *S. cerevisiae* mating factor or ii) pro-region of *P. pastoris* Epxl. Accordingly, described herein are isolated nucleic acids encoding such secretion signal peptides, in some cases linked to a recombinant protein such as a human milk protein, as well as cells comprising such nucleic acids and methods for producing and collecting recombinant proteins from such cells.

**[0010]** Described herein, in some embodiments, is an isolated nucleic acid encoding a polypeptide comprising a

sequence having at least 90% sequence identity to SEQ ID NO:1, 2, 3, or 4. In some embodiments, the sequence comprises SEQ ID NO:1, 2, 3, or 4. In some embodiments, the polypeptide further comprises a sequence of a mammalian protein. In some embodiments, the mammalian protein is a human milk protein. In some embodiments, the human milk protein is secretory IgA (sIgA), xanthine dehydrogenase, lactoferrin, lactoperoxidase, butyrophilin, lactadherin, adiponectin,  $\beta$ -casein,  $\kappa$ -casein, leptin, lysozyme, or  $\alpha$ -lactalbumin. In some embodiments, the human milk protein is human lactoferrin.

**[0011]** In some embodiments, the sequence has at least 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% sequence identity to SEQ ID NO:1. In some embodiments, the sequence comprises SEQ ID NO:1. In some embodiments, the isolated nucleic acid comprises a nucleic acid sequence having at least 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% identity to SEQ ID NO:41. In some embodiments, the nucleic acid sequence comprises SEQ ID NO:41. In some embodiments, the polypeptide comprises SEQ ID NO:5. In some embodiments, the isolated nucleic acid comprises a nucleic acid sequence having at least 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% identity to SEQ ID NO:46. In some embodiments, the nucleic acid sequence comprises SEQ ID NO:46.

**[0012]** In some embodiments, the sequence has at least 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% sequence identity to SEQ ID NO:2. In some embodiments, the sequence comprises SEQ ID NO:2. In some embodiments, the isolated nucleic acid comprises a nucleic acid sequence having at least 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% identity to SEQ ID NO:42. In some embodiments, the nucleic acid sequence comprises SEQ ID NO:42. In some embodiments, the polypeptide comprises SEQ ID NO:6. In some embodiments, the isolated nucleic acid comprises a nucleic acid sequence having at least 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% identity to SEQ ID NO:47. In some embodiments, the nucleic acid sequence comprises SEQ ID NO:47.

**[0013]** In some embodiments, the sequence has at least 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% sequence identity to SEQ ID NO:3. In some embodiments, the sequence comprises SEQ ID NO:3. In some embodiments, the isolated nucleic acid comprises a nucleic acid sequence having at least 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% identity to SEQ ID NO:43. In some embodiments, the nucleic acid sequence comprises SEQ ID NO:43. In some embodiments, the polypeptide comprises SEQ ID NO:7. In some embodiments, the isolated nucleic acid comprises a nucleic acid sequence having at least 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% identity to SEQ ID NO:48. In some embodiments, the nucleic acid sequence comprises SEQ ID NO:48.

**[0014]** In some embodiments, the sequence has at least 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% sequence identity to SEQ ID NO:4. In some embodiments, the sequence comprises SEQ ID NO:4. In some embodiments, the isolated nucleic acid comprises a nucleic acid sequence having at least 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% identity to SEQ ID NO:44. In some embodiments, the nucleic acid sequence comprises SEQ ID NO:44. In some embodiments, the polypeptide comprises SEQ ID NO:8. In some embodiments, the isolated nucleic acid comprises a nucleic acid sequence having at least 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% identity to SEQ ID NO:49. In some embodiments, the nucleic acid sequence comprises SEQ ID NO:49.

**[0015]** Also disclosed herein, in some embodiments, is a vector comprising a nucleic acid (e.g., an isolated nucleic acid or sequence or portion thereof) disclosed herein.

**[0016]** Further disclosed, in some aspects, is an engineered eukaryotic cell comprising a nucleic acid disclosed herein. In some embodiments, the cell is a fungal cell. In some embodiments, the fungal cell is a *Arxula*, *Aspegillus*, *Aurantiochytrium*, *Candida*, *Claviceps*, *Cryptococcus*, *Cunninghamella*, *Geotrichum*, *Hansenula*, *Kluyveromyces*, *Kodamaea*, *Komagataella*, *Leucosporidiella*, *Lipomyces*, *Mortierella*, *Ogataea*, *Pichia*, *Prototheca*, *Rhizopus*, *Rhodospiridium*, *Rhodotorula*, *Saccharomyces*, *Schizosaccharomyces*, *Tremella*, *Trichosporon*, *Wickerhamomyces*, or *Yarrowia* cell. In some embodiments, the cell is a yeast cell. In some embodiments, the yeast cell is a *Komagataella* cell. In some embodiments, the yeast cell is a *Komagataella phaffii*, *Komagataella pastoris*, or *Komagataella pseudopastoris* cell. In some aspects, the nucleic acid is integrated into the genome of the cell. In some aspects, the nucleic acid is not integrated into the genome of the cell.

**[0017]** Also disclosed, in some aspects, is a method for producing a secreted protein, the method comprising growing an engineered eukaryotic cell of the present disclosure under conditions sufficient to secrete the polypeptide from the cell. In some embodiments, the method further comprises collecting the secreted protein. In some aspects, the secreted protein is a human milk protein. In some embodiments, the human milk protein is secretory IgA (sIgA), xanthine dehydrogenase, lactoferrin, lactoperoxidase, butyrophilin, lactadherin, adiponectin,  $\beta$ -casein,  $\kappa$ -casein, leptin, lysozyme, or  $\alpha$ -lactalbumin. In some embodiments, the human milk protein is human lactoferrin. In some embodiments, the human milk protein comprises one or more human-like N-glycans. In some embodiments, the method further comprises generating a mixture comprising the human milk protein and one or more components of an infant formula.

**[0018]** Further disclosed herein, in some aspects, is an engineered yeast cell comprising a nucleic acid encoding a

polypeptide comprising a sequence having at least 90% sequence identity to SEQ ID NO:1, 2, 3, or 4. In some embodiments, the sequence comprises SEQ ID NO:1, 2, 3, or 4. In some embodiments, the sequence comprises SEQ ID NO:1. In some embodiments, the sequence comprises SEQ ID NO:2. In some embodiments, the sequence comprises SEQ ID NO:3. In some embodiments, the sequence comprises SEQ ID NO:4. In some embodiments, the polypeptide further comprises a sequence of a mammalian protein. In some embodiments, the mammalian protein is a human milk protein. In some embodiments, the human milk protein is secretory IgA (sIgA), xanthine dehydrogenase, lactoferrin, lactoperoxidase, butyrophilin, lactadherin, adiponectin,  $\beta$ -casein,  $\kappa$ -casein, leptin, lysozyme, or  $\alpha$ -lactalbumin. In some embodiments, the human milk protein is human lactoferrin.

**[0019]** Described herein, in some aspects, is an engineered yeast cell comprising: (a) a first nucleic acid encoding a polypeptide comprising: (i) a sequence having at least 90% sequence identity to SEQ ID NO:1, 2, 3, or 4 and (ii) a sequence of a human milk protein; and (b) a second nucleic acid encoding an  $\alpha$ -1,2-mannosidase (Man-I) protein, wherein the cell does not express a functional OCH1 protein. In some embodiments, the sequence of (i) comprises SEQ ID NO:1, 2, 3, or 4. In some embodiments, the human milk protein is secretory IgA (sIgA), xanthine dehydrogenase, lactoferrin, lactoperoxidase, butyrophilin, lactadherin, adiponectin,  $\beta$ -casein,  $\kappa$ -casein, leptin, lysozyme, or  $\alpha$ -lactalbumin. In some embodiments, the human milk protein is human lactoferrin. In some embodiments, the human milk protein is human  $\alpha$ -lactalbumin. In some embodiments, the Man-I protein is fused to a HDEL C-terminal tag. In some embodiments, the cell further comprises a third nucleic acid encoding one or more of: (a) a N-acetylglucosaminyltransferase I (GnT-I) protein; (b) an  $\alpha$ -1,3/6-Mannosidase (Man-II) protein; (c) a  $\beta$ -1,2-acetylglucosaminyltransferase (GnT-II) protein; and (d) a  $\beta$ -1,4-galactosyltransferase (GalT) protein. In some embodiments, the yeast cell is a *Komagataella phaffii*, *Komagataella pastoris*, or *Komagataella pseudopastoris* cell. In some aspects, the nucleic acid is integrated into the genome of the cell. In some aspects, the nucleic acid is not integrated into the genome of the cell.

**[0020]** It is contemplated that any embodiment discussed in this specification can be implemented with respect to any method or composition of the disclosed embodiments, and vice versa. Furthermore, compositions of the embodiments disclosed herein can be used to achieve methods of those embodiments.

**[0021]** Other objects, features and advantages of the present embodiments disclosed herein will become apparent from the following detailed description. It should be understood, however, that the detailed description and the specific examples, while indicating specific embodiments, are given by way of illustration only, since various changes and modifications within the spirit and scope of the embodiments disclosed herein will become apparent to those skilled in the art from this detailed description.

## BRIEF DESCRIPTION OF THE DRAWINGS

**[0022]** The following drawings form part of the present specification and are included to further demonstrate certain aspects of the present disclosure. This may be better understood by reference to one or more of these drawings in combination with the detailed description of specific embodiments presented herein.

**FIG. 1** is an image of a Western Blot of supernatants. Lane 1 is loaded with a protein standard, Genscript, M00624 (ThermoFisher Scientific, Waltham, MA, USA). Lane 2 is loaded with lactoferrin from Human Milk, Sigma Aldrich, SRP6519 (Sigma Aldrich, St. Louis, MO, USA). Lane 3 is loaded with a control (*Saccharomyces cerevisiae* pre-pro-MF $\alpha$ ). Lane 4 is loaded with the negative control, a supernatant of untransformed yeast cells. Lanes 5-6 are loaded with supernatant from SP2-lactoferrin transformed yeast cells. Lanes 7-8 are loaded with supernatant from SP3-lactoferrin transformed yeast cells. Lanes 9-10 are loaded with SP1-lactoferrin transformed yeast cells.

**FIG. 2** is a bar graph showing protein expression levels. Quantification of extracellular protein was performed via ELISA.

## DETAILED DESCRIPTION

**[0023]** Described herein is the generation of novel synthetic secretion signal peptides. Also disclosed are cells (e.g., fungal cells such as yeast cells) engineered to express one or more exogenous proteins (e.g., human milk proteins) comprising such signal peptides. As disclosed herein, the in-frame fusion of "pre-region" sequences from *P. pastoris* Ost1 or Pst1 and "pro-region" sequences from *S. cerevisiae* mating factor  $\alpha$  or *P. pastoris* Epx1 can facilitate increased extracellular protein production compared with previously used signal peptides. The disclosed signal peptides include, for example, peptides comprising SEQ ID NOs:1, 2, 3, or 4, as well as peptides comprising 1, 2, 3, 4, or 5 amino acid substitutions (or more) relative to SEQ ID NO: 1, 2, 3, or 4. As described herein, in-frame fusion of these hybrid signal peptides to the N-terminus of mammalian proteins (e.g., human milk proteins such as lactoferrin or  $\alpha$ -lactalbumin) promotes highly efficient protein secretion.

## I. Definitions

**[0024]** The term "biologically-active portion" refers to an amino acid sequence that is less than a full-length amino acid sequence, but exhibits at least one activity of the full length sequence. For example, a biologically-active portion of an enzyme may refer to one or more domains of an enzyme having the catalytic activity of the enzyme (i.e., may be a catalytic domain). In some aspects, a biologically-active portion of an enzyme is a portion of the enzyme comprising a catalytic domain of the enzyme. Biologically-active portions of a protein include peptides or polypeptides comprising amino acid sequences sufficiently identical to or derived from the amino acid sequence of the protein, which include fewer amino acids than the full length protein, and exhibit at least one activity (e.g., enzymatic activity, functional activity, etc.) of the protein.

**[0025]** The term "exogenous" refers to anything that is introduced into a cell or has been introduced into a cell. An "exogenous nucleic acid" is a nucleic acid that enters or has entered a cell through the cell membrane. An "exogenous nucleic acid sequence" is a nucleic acid sequence of an exogenous nucleic acid. An exogenous nucleic acid may contain a nucleotide sequence that exists in the native genome of a cell and/or nucleotide sequences that did not previously exist in the cell's genome. Exogenous nucleic acids include exogenous genes. An "exogenous gene" is a nucleic acid that codes for the expression of an RNA and/or protein that has been introduced into a cell (e.g., by transformation/transfection), and is also referred to as a "transgene." A cell comprising an exogenous nucleic acid may be referred to as a recombinant cell, into which additional exogenous gene(s) may be introduced. The exogenous gene may be from the same or different species relative to the cell being transformed. Thus, an exogenous gene can include a native gene that occupies a different location in the genome of the cell or is under different control, relative to the endogenous copy of the gene. An exogenous gene may be present in more than one copy in the cell. An exogenous gene may be maintained in a cell as an insertion into the genome (nuclear, mitochondrial, or plastid) or as an episomal molecule.

**[0026]** "In operable linkage" (or "operably linked") refers to a functional linkage between two nucleic acid sequences, such a control sequence (typically a promoter) and the linked sequence (typically a sequence that encodes a protein, also called a coding sequence). A promoter is in operable linkage with a gene if it can mediate transcription of the gene.

**[0027]** The term "native" refers to the composition of a cell or parent cell prior to a transformation event. A "native gene" (also "endogenous gene") refers to a nucleotide sequence that encodes a protein that has not been introduced into a cell by a transformation event. A "native protein" (also "endogenous protein") refers to an amino acid sequence that is encoded by a native gene.

**[0028]** "Recombinant" refers to a cell, nucleic acid, protein, or vector, which has been modified due to introduction of an exogenous nucleic acid or alteration of a native nucleic acid. Resulting cells, nucleic acids, proteins or vectors are considered recombinant, as are progeny, offspring, duplications or replications of these are also considered recombinant. Thus, e.g., recombinant cells can express genes that are not found within the native (non-recombinant) form of the cell or express native genes differently than those same genes are expressed by a non-recombinant cell. Recombinant cells can, without limitation, include recombinant nucleic acids that encode for a gene product or for suppression elements such as mutations, knockouts, antisense, interfering RNA (RNAi), or dsRNA that reduce the levels of active gene product in a cell. A "recombinant nucleic acid" is derived from nucleic acid originally formed in vitro, in general, by the manipulation of nucleic acid, e.g., using polymerases, ligases, exonucleases, and endonucleases, or otherwise is in a form not normally found in nature. Once a recombinant nucleic acid is made and introduced into a host cell or organism, it may replicate using the in vivo cellular machinery of the host cell; however, such nucleic acids, once produced recombinantly, although subsequently replicated intracellularly, are still considered recombinant for purposes of this disclosure. Additionally, a recombinant nucleic acid refers to nucleotide sequences that comprise an endogenous nucleotide sequence and an exogenous nucleotide sequence; thus, an endogenous gene that has undergone recombination with an exogenous promoter is a recombinant nucleic acid. A "recombinant protein" is a protein made using recombinant techniques, i.e., through the expression of a recombinant nucleic acid.

**[0029]** "Transformation" refers to the transfer of a nucleic acid into a host organism or the genome of a host organism. Host organisms (and their progeny) containing the transformed nucleic acid fragments are referred to as "recombinant", "transgenic" or "transformed" organisms. Thus, isolated polynucleotides of the present disclosure can be incorporated into recombinant constructs, typically DNA constructs, capable of introduction into and replication in a host cell. Such a construct can be a vector that includes a replication system and sequences that are capable of transcription and translation of a polypeptide-encoding sequence in a given host cell. Typically, expression vectors include, for example, one or more cloned genes under the transcriptional control of 5' and 3' regulatory sequences and a selectable marker. Such vectors also can contain a promoter regulatory region (e.g., a regulatory region controlling inducible or constitutive, environmentally- or developmentally-regulated, or location-specific expression), a transcription initiation start site, a ribosome binding site, a transcription termination site, and/or a polyadenylation signal. Alternatively, a cell may be transformed with a single genetic element, such as a promoter, which may result in genetically stable inheritance upon integrating into the host organism's genome, such as by homologous recombination.

**[0030]** The term "transformed cell" refers to a cell that has undergone a transformation. Thus, a transformed cell

comprises the parent's genome and an inheritable genetic modification. Embodiments include progeny and offspring of such transformed cells.

**[0031]** The term "vector" refers to the means by which a nucleic acid can be propagated and/or transferred between organisms, cells, or cellular components. Vectors include plasmids, linear DNA fragments, viruses, bacteriophage, proviruses, phagemids, transposons, and artificial chromosomes, and the like, that may or may not be able to replicate autonomously or integrate into a chromosome of a host cell.

**[0032]** "Individual," "subject," and "patient" are used interchangeably and can refer to a human or non-human.

**[0033]** Throughout this application, the term "about" is used to indicate that a value includes the inherent variation of error for the measurement or quantitation method.

**[0034]** The use of the word "a" or "an" when used in conjunction with the term "comprising" may mean "one," but it is also consistent with the meaning of "one or more," "at least one," and "one or more than one."

**[0035]** The phrase "and/or" means "and" or "or". To illustrate, A, B, and/or C includes: A alone, B alone, C alone, a combination of A and B, a combination of A and C, a combination of B and C, or a combination of A, B, and C. In other words, "and/or" operates as an inclusive or.

**[0036]** The words "comprising" (and any form of comprising, such as "comprise" and "comprises"), "having" (and any form of having, such as "have" and "has"), "including" (and any form of including, such as "includes" and "include") or "containing" (and any form of containing, such as "contains" and "contain") are inclusive or open-ended and do not exclude additional, unrecited elements or method steps.

**[0037]** The compositions and methods for their use can "comprise," "consist essentially of," or "consist of" any of the ingredients or steps disclosed throughout the specification. Compositions and methods "consisting essentially of" any of the ingredients or steps disclosed limits the scope of the claim to the specified materials or steps which do not materially affect the basic and novel characteristic of the claimed embodiment.

## II. Proteins and Nucleic Acids

**[0038]** As used herein, a "protein" or "polypeptide" refers to a molecule comprising at least five amino acid residues. As used herein, the term "wild-type" refers to the endogenous version of a molecule that occurs naturally in an organism. In some embodiments, wild-type versions of a protein or polypeptide are employed, however, in many embodiments of the disclosure, a modified protein or polypeptide is employed. The terms described above may be used interchangeably.

A "modified protein" or "modified polypeptide" or a "variant" refers to a protein or polypeptide whose chemical structure, particularly its amino acid sequence, is altered with respect to the wild-type protein or polypeptide. In some embodiments, a modified/variant protein or polypeptide has at least one modified activity or function (recognizing that proteins or polypeptides may have multiple activities or functions). It is specifically contemplated that a modified/variant protein or polypeptide may be altered with respect to one activity or function yet retain a wild-type activity or function in other respects.

**[0039]** Where a protein is specifically mentioned herein, it is in general a reference to a native (wild-type) or recombinant (modified) protein or, optionally, a protein in which any signal sequence has been removed. The protein may be isolated directly from the organism of which it is native, produced by recombinant DNA/exogenous expression methods, or produced by solid-phase peptide synthesis (SPPS) or other in vitro methods. In particular embodiments, there are isolated nucleic acid segments and recombinant vectors incorporating nucleic acid sequences that encode a polypeptide. The term "recombinant" may be used in conjunction with a polypeptide or the name of a specific polypeptide, and this generally refers to a polypeptide produced from a nucleic acid molecule that has been manipulated in vitro or that is a replication product of such a molecule.

**[0040]** In certain embodiments the size of a protein or polypeptide (wild-type or modified) may comprise, but is not limited to, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200, 210, 220, 230, 240, 250, 275, 300, 325, 350, 375, 400, 425, 450, 475, 500, 525, 550, 575, 600, 625, 650, 675, 700, 725, 750, 775, 800, 825, 850, 875, 900, 925, 950, 975, 1000, 1100, 1200, 1300, 1400, 1500, 1750, 2000, 2250, 2500 amino acid residues or greater, and any range derivable therein, or derivative of a corresponding amino sequence described or referenced herein. It is contemplated that polypeptides may be mutated by truncation, rendering them shorter than their corresponding wild-type form, also, they might be altered by fusing or conjugating a heterologous protein or polypeptide sequence with a particular function (e.g., for targeting or localization, for enhanced immunogenicity, for purification purposes, etc.). As used herein, the term "domain" refers to any distinct functional or structural unit of a protein or polypeptide, and generally refers to a sequence of amino acids with a structure or function recognizable by one skilled in the art.

**[0041]** The term "polynucleotide" refers to a nucleic acid molecule that either is recombinant or has been isolated from total genomic nucleic acid. Included within the term "polynucleotide" are oligonucleotides (nucleic acids 100 residues or less in length), recombinant vectors, including, for example, plasmids, cosmids, phage, viruses, and the like. Polynu-

cleotides include, in certain aspects, regulatory sequences, isolated substantially away from their naturally occurring genes or protein encoding sequences. Polynucleotides may be single- stranded (coding or antisense) or double- stranded, and may be RNA, DNA (genomic, cDNA or synthetic), analogs thereof, or a combination thereof. Additional coding or noncoding sequences may, but need not, be present within a polynucleotide.

**[0042]** In this respect, the term "gene," "polynucleotide," or "nucleic acid" is used to refer to a nucleic acid that encodes a protein, polypeptide, or peptide (including any sequences required for proper transcription, post-translational modification, or localization). As will be understood by those in the art, this term encompasses genomic sequences, expression cassettes, cDNA sequences, and smaller engineered nucleic acid segments that express, or may be adapted to express, proteins, polypeptides, domains, peptides, fusion proteins, and mutants. A nucleic acid encoding all or part of a polypeptide may contain a contiguous nucleic acid sequence encoding all or a portion of such a polypeptide. It also is contemplated that a particular polypeptide may be encoded by nucleic acids containing variations having slightly different nucleic acid sequences but, nonetheless, encode the same or substantially similar protein.

**[0043]** In certain embodiments, there are polynucleotide variants having substantial identity to the sequences disclosed herein; those comprising at least 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, or 99% or higher sequence identity, including all values and ranges there between, compared to a polynucleotide sequence provided herein using the methods described herein (e.g., BLAST analysis using standard parameters). In certain aspects, the isolated polynucleotide will comprise a nucleotide sequence encoding a polypeptide that has at least 90%, and in some cases 95% and above, identity to an amino acid sequence described herein, over the entire length of the sequence; or a nucleotide sequence complementary to said isolated polynucleotide.

**[0044]** The nucleic acid segments, regardless of the length of the coding sequence itself, may be combined with other nucleic acid sequences, such as promoters, polyadenylation signals, additional restriction enzyme sites, multiple cloning sites, other coding segments, and the like, such that their overall length may vary considerably. The nucleic acids can be any length. They can be, for example, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 75, 100, 125, 175, 200, 250, 300, 350, 400, 450, 500, 750, 1000, 1500, 3000, 5000 or more nucleotides in length, and/or can comprise one or more additional sequences, for example, regulatory sequences, and/or be a part of a larger nucleic acid, for example, a vector. It is therefore contemplated that a nucleic acid fragment of almost any length may be employed, with the total length preferably being limited by the ease of preparation and use in the intended recombinant nucleic acid protocol. In some cases, a nucleic acid sequence may encode a polypeptide sequence with additional heterologous coding sequences, for example to allow for purification of the polypeptide, transport, secretion, post-translational modification, or for therapeutic benefits such as targeting or efficacy. As discussed above, a tag or other heterologous polypeptide may be added to the modified polypeptide-encoding sequence, wherein "heterologous" refers to a polypeptide that is not the same as the modified polypeptide.

**[0045]** The polypeptides, proteins, or polynucleotides encoding such polypeptides or proteins of the disclosure may include 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, or 50 (or any derivable range therein) or more variant amino acids or nucleic acid substitutions or be at least 60%, 61%, 62%, 63%, 64%, 65%, 66%, 67%, 68%, 69%, 70%, 71%, 72%, 73%, 74%, 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, or 100% (or any derivable range therein) similar, identical, or homologous with at least, or at most 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193, 194, 195, 196, 197, 198, 199, 200, 201, 202, 203, 204, 205, 206, 207, 208, 209, 210, 211, 212, 213, 214, 215, 216, 217, 218, 219, 220, 221, 222, 223, 224, 225, 226, 227, 228, 229, 230, 231, 232, 233, 234, 235, 236, 237, 238, 239, 240, 241, 242, 243, 244, 245, 246, 247, 248, 249, 250, 300, 400, 500, 550, 1000 or more contiguous amino acids or nucleic acids, or any range derivable therein, of SEQ ID NOs: 1-49.

**[0046]** In some embodiments, the protein or polypeptide may comprise amino acids 1 to 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193, 194, 195, 196, 197, 198, 199, 200, 201, 202, 203, 204, 205, 206, 207, 208, 209, 210, 211, 212, 213, 214, 215, 216, 217, 218, 219, 220, 221, 222, 223,

224, 225, 226, 227, 228, 229, 230, 231, 232, 233, 234, 235, 236, 237, 238, 239, 240, 241, 242, 243, 244, 245, 246, 247, 248, 249, 250, 251, 252, 253, 254, 255, 256, 257, 258, 259, 260, 261, 262, 263, 264, 265, 266, 267, 268, 269, 270, 271, 272, 273, 274, 275, 276, 277, 278, 279, 280, 281, 282, 283, 284, 285, 286, 287, 288, 289, 290, 291, 292, 293, 294, 295, 296, 297, 298, 299, 300, 301, 302, 303, 304, 305, 306, 307, 308, 309, 310, 311, 312, 313, 314, 315, 316, 317, 318, 319, 320, 321, 322, 323, 324, 325, 326, 327, 328, 329, 330, 331, 332, 333, 334, 335, 336, 337, 338, 339, 340, 341, 342, 343, 344, 345, 346, 347, 348, 349, 350, 351, 352, 353, 354, 355, 356, 357, 358, 359, 360, 361, 362, 363, 364, 365, 366, 367, 368, 369, 370, 371, 372, 373, 374, 375, 376, 377, 378, 379, 380, 381, 382, 383, 384, 385, 386, 387, 388, 389, 390, 391, 392, 393, 394, 395, 396, 397, 398, 399, 400, 401, 402, 403, 404, 405, 406, 407, 408, 409, 410, 411, 412, 413, 414, 415, 416, 417, 418, 419, 420, 421, 422, 423, 424, 425, 426, 427, 428, 429, 430, 431, 432, 433, 434, 435, 436, 437, 438, 439, 440, 441, 442, 443, 444, 445, 446, 447, 448, 449, 450, 451, 452, 453, 454, 455, 456, 457, 458, 459, 460, 461, 462, 463, 464, 465, 466, 467, 468, 469, 470, 471, 472, 473, 474, 475, 476, 477, 478, 479, 480, 481, 482, 483, 484, 485, 486, 487, 488, 489, 490, 491, 492, 493, 494, 495, 496, 497, 498, 499, 500, 501, 502, 503, 504, 505, 506, 507, 508, 509, 510, 511, 512, 513, 514, 515, 516, 517, 518, 519, 520, 521, 522, 523, 524, 525, 526, 527, 528, 529, 530, 531, 532, 533, 534, 535, 536, 537, 538, 539, 540, 541, 542, 543, 544, 545, 546, 547, 548, 549, 550, 551, 552, 553, 554, 555, 556, 557, 558, 559, 560, 561, 562, 563, 564, 565, 566, 567, 568, 569, 570, 571, 572, 573, 574, 575, 576, 577, 578, 579, 580, 581, 582, 583, 584, 585, 586, 587, 588, 589, 590, 591, 592, 593, 594, 595, 596, 597, 598, 599, 600, 601, 602, 603, 604, 605, 606, 607, 608, 609, 610, 611, 612, 613, 614, 615, 616, 617, 618, 619, 620, 621, 622, 623, 624, 625, 626, 627, 628, 629, 630, 631, 632, 633, 634, 635, 636, 637, 638, 639, 640, 641, 642, 643, 644, 645, 646, 647, 648, 649, 650, 651, 652, 653, 654, 655, 656, 657, 658, 659, 660, 661, 662, 663, 664, 665, 666, 667, 668, 669, 670, 671, 672, 673, 674, 675, 676, 677, 678, 679, 680, 681, 682, 683, 684, 685, 686, 687, 688, 689, 690, 691, 692, 693, 694, 695, 696, 697, 698, 699, or 700 (or any derivable range therein) of SEQ ID NOs:1-14 or 34-40.

**[0047]** In some embodiments, the protein, polypeptide, or nucleic acid may comprise 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193, 194, 195, 196, 197, 198, 199, 200, 201, 202, 203, 204, 205, 206, 207, 208, 209, 210, 211, 212, 213, 214, 215, 216, 217, 218, 219, 220, 221, 222, 223, 224, 225, 226, 227, 228, 229, 230, 231, 232, 233, 234, 235, 236, 237, 238, 239, 240, 241, 242, 243, 244, 245, 246, 247, 248, 249, 250, 251, 252, 253, 254, 255, 256, 257, 258, 259, 260, 261, 262, 263, 264, 265, 266, 267, 268, 269, 270, 271, 272, 273, 274, 275, 276, 277, 278, 279, 280, 281, 282, 283, 284, 285, 286, 287, 288, 289, 290, 291, 292, 293, 294, 295, 296, 297, 298, 299, 300, 301, 302, 303, 304, 305, 306, 307, 308, 309, 310, 311, 312, 313, 314, 315, 316, 317, 318, 319, 320, 321, 322, 323, 324, 325, 326, 327, 328, 329, 330, 331, 332, 333, 334, 335, 336, 337, 338, 339, 340, 341, 342, 343, 344, 345, 346, 347, 348, 349, 350, 351, 352, 353, 354, 355, 356, 357, 358, 359, 360, 361, 362, 363, 364, 365, 366, 367, 368, 369, 370, 371, 372, 373, 374, 375, 376, 377, 378, 379, 380, 381, 382, 383, 384, 385, 386, 387, 388, 389, 390, 391, 392, 393, 394, 395, 396, 397, 398, 399, 400, 401, 402, 403, 404, 405, 406, 407, 408, 409, 410, 411, 412, 413, 414, 415, 416, 417, 418, 419, 420, 421, 422, 423, 424, 425, 426, 427, 428, 429, 430, 431, 432, 433, 434, 435, 436, 437, 438, 439, 440, 441, 442, 443, 444, 445, 446, 447, 448, 449, 450, 451, 452, 453, 454, 455, 456, 457, 458, 459, 460, 461, 462, 463, 464, 465, 466, 467, 468, 469, 470, 471, 472, 473, 474, 475, 476, 477, 478, 479, 480, 481, 482, 483, 484, 485, 486, 487, 488, 489, 490, 491, 492, 493, 494, 495, 496, 497, 498, 499, 500, 501, 502, 503, 504, 505, 506, 507, 508, 509, 510, 511, 512, 513, 514, 515, 516, 517, 518, 519, 520, 521, 522, 523, 524, 525, 526, 527, 528, 529, 530, 531, 532, 533, 534, 535, 536, 537, 538, 539, 540, 541, 542, 543, 544, 545, 546, 547, 548, 549, 550, 551, 552, 553, 554, 555, 556, 557, 558, 559, 560, 561, 562, 563, 564, 565, 566, 567, 568, 569, 570, 571, 572, 573, 574, 575, 576, 577, 578, 579, 580, 581, 582, 583, 584, 585, 586, 587, 588, 589, 590, 591, 592, 593, 594, 595, 596, 597, 598, 599, 600, 601, 602, 603, 604, 605, 606, 607, 608, 609, 610, 611, 612, 613, 614, 615, 616, 617, 618, 619, 620, 621, 622, 623, 624, 625, 626, 627, 628, 629, 630, 631, 632, 633, 634, 635, 636, 637, 638, 639, 640, 641, 642, 643, 644, 645, 646, 647, 648, 649, 650, 651, 652, 653, 654, 655, 656, 657, 658, 659, 660, 661, 662, 663, 664, 665, 666, 667, 668, 669, 670, 671, 672, 673, 674, 675, 676, 677, 678, 679, 680, 681, 682, 683, 684, 685, 686, 687, 688, 689, 690, 691, 692, 693, 694, 695, 696, 697, 698, 699, or 700 (or any derivable range therein) contiguous amino acids of SEQ ID NOs:1-49.

**[0048]** In some embodiments, the polypeptide, protein, or nucleic acid may comprise at least, at most, or exactly 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169,



170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193, 194, 195, 196, 197, 198, 199, 200, 201, 202, 203, 204, 205, 206, 207, 208, 209, 210, 211, 212, 213, 214, 215, 216, 217, 218, 219, 220, 221, 222, 223, 224, 225, 226, 227, 228, 229, 230, 231, 232, 233, 234, 235, 236, 237, 238, 239, 240, 241, 242, 243, 244, 245, 246, 247, 248, 249, 250, 251, 252, 253, 254, 255, 256, 257, 258, 259, 260, 261, 262, 263, 264, 265, 266, 267, 268, 269, 270, 271, 272, 273, 274, 275, 276, 277, 278, 279, 280, 281, 282, 283, 284, 285, 286, 287, 288, 289, 290, 291, 292, 293, 294, 295, 296, 297, 298, 299, 300, 301, 302, 303, 304, 305, 306, 307, 308, 309, 310, 311, 312, 313, 314, 315, 316, 317, 318, 319, 320, 321, 322, 323, 324, 325, 326, 327, 328, 329, 330, 331, 332, 333, 334, 335, 336, 337, 338, 339, 340, 341, 342, 343, 344, 345, 346, 347, 348, 349, 350, 351, 352, 353, 354, 355, 356, 357, 358, 359, 360, 361, 362, 363, 364, 365, 366, 367, 368, 369, 370, 371, 372, 373, 374, 375, 376, 377, 378, 379, 380, 381, 382, 383, 384, 385, 386, 387, 388, 389, 390, 391, 392, 393, 394, 395, 396, 397, 398, 399, 400, 401, 402, 403, 404, 405, 406, 407, 408, 409, 410, 411, 412, 413, 414, 415, 416, 417, 418, 419, 420, 421, 422, 423, 424, 425, 426, 427, 428, 429, 430, 431, 432, 433, 434, 435, 436, 437, 438, 439, 440, 441, 442, 443, 444, 445, 446, 447, 448, 449, 450, 451, 452, 453, 454, 455, 456, 457, 458, 459, 460, 461, 462, 463, 464, 465, 466, 467, 468, 469, 470, 471, 472, 473, 474, 475, 476, 477, 478, 479, 480, 481, 482, 483, 484, 485, 486, 487, 488, 489, 490, 491, 492, 493, 494, 495, 496, 497, 498, 499, 500, 501, 502, 503, 504, 505, 506, 507, 508, 509, 510, 511, 512, 513, 514, 515, 516, 517, 518, 519, 520, 521, 522, 523, 524, 525, 526, 527, 528, 529, 530, 531, 532, 533, 534, 535, 536, 537, 538, 539, 540, 541, 542, 543, 544, 545, 546, 547, 548, 549, 550, 551, 552, 553, 554, 555, 556, 557, 558, 559, 560, 561, 562, 563, 564, 565, 566, 567, 568, 569, 570, 571, 572, 573, 574, 575, 576, 577, 578, 579, 580, 581, 582, 583, 584, 585, 586, 587, 588, 589, 590, 591, 592, 593, 594, 595, 596, 597, 598, 599, 600, 601, 602, 603, 604, 605, 606, 607, 608, 609, 610, 611, 612, 613, 614, 615, 616, 617, 618, 619, 620, 621, 622, 623, 624, 625, 626, 627, 628, 629, 630, 631, 632, 633, 634, 635, 636, 637, 638, 639, 640, 641, 642, 643, 644, 645, 646, 647, 648, 649, 650, 651, 652, 653, 654, 655, 656, 657, 658, 659, 660, 661, 662, 663, 664, 665, 666, 667, 668, 669, 670, 671, 672, 673, 674, 675, 676, 677, 678, 679, 680, 681, 682, 683, 684, 685, 686, 687, 688, 689, 690, 691, 692, 693, 694, 695, 696, 697, 698, 699, or 700 (or any derivable range therein) contiguous amino acids of SEQ ID NOs:1-49 that are at least, at most, or exactly 60%, 61%, 62%, 63%, 64%, 65%, 66%, 67%, 68%, 69%, 70%, 71%, 72%, 73%, 74%, 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, or 100% (or any derivable range therein) similar, identical, or homologous with one of SEQ ID NOs:1-49.

**[0049]** In some aspects there is a nucleic acid molecule or polypeptide starting at position 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193, 194, 195, 196, 197, 198, 199, 200, 201, 202, 203, 204, 205, 206, 207, 208, 209, 210, 211, 212, 213, 214, 215, 216, 217, 218, 219, 220, 221, 222, 223, 224, 225, 226, 227, 228, 229, 230, 231, 232, 233, 234, 235, 236, 237, 238, 239, 240, 241, 242, 243, 244, 245, 246, 247, 248, 249, 250, 251, 252, 253, 254, 255, 256, 257, 258, 259, 260, 261, 262, 263, 264, 265, 266, 267, 268, 269, 270, 271, 272, 273, 274, 275, 276, 277, 278, 279, 280, 281, 282, 283, 284, 285, 286, 287, 288, 289, 290, 291, 292, 293, 294, 295, 296, 297, 298, 299, 300, 301, 302, 303, 304, 305, 306, 307, 308, 309, 310, 311, 312, 313, 314, 315, 316, 317, 318, 319, 320, 321, 322, 323, 324, 325, 326, 327, 328, 329, 330, 331, 332, 333, 334, 335, 336, 337, 338, 339, 340, 341, 342, 343, 344, 345, 346, 347, 348, 349, 350, 351, 352, 353, 354, 355, 356, 357, 358, 359, 360, 361, 362, 363, 364, 365, 366, 367, 368, 369, 370, 371, 372, 373, 374, 375, 376, 377, 378, 379, 380, 381, 382, 383, 384, 385, 386, 387, 388, 389, 390, 391, 392, 393, 394, 395, 396, 397, 398, 399, 400, 401, 402, 403, 404, 405, 406, 407, 408, 409, 410, 411, 412, 413, 414, 415, 416, 417, 418, 419, 420, 421, 422, 423, 424, 425, 426, 427, 428, 429, 430, 431, 432, 433, 434, 435, 436, 437, 438, 439, 440, 441, 442, 443, 444, 445, 446, 447, 448, 449, 450, 451, 452, 453, 454, 455, 456, 457, 458, 459, 460, 461, 462, 463, 464, 465, 466, 467, 468, 469, 470, 471, 472, 473, 474, 475, 476, 477, 478, 479, 480, 481, 482, 483, 484, 485, 486, 487, 488, 489, 490, 491, 492, 493, 494, 495, 496, 497, 498, 499, 500, 501, 502, 503, 504, 505, 506, 507, 508, 509, 510, 511, 512, 513, 514, 515, 516, 517, 518, 519, 520, 521, 522, 523, 524, 525, 526, 527, 528, 529, 530, 531, 532, 533, 534, 535, 536, 537, 538, 539, 540, 541, 542, 543, 544, 545, 546, 547, 548, 549, 550, 551, 552, 553, 554, 555, 556, 557, 558, 559, 560, 561, 562, 563, 564, 565, 566, 567, 568, 569, 570, 571, 572, 573, 574, 575, 576, 577, 578, 579, 580, 581, 582, 583, 584, 585, 586, 587, 588, 589, 590, 591, 592, 593, 594, 595, 596, 597, 598, 599, 600, 601, 602, 603, 604, 605, 606, 607, 608, 609, 610, 611, 612, 613, 614, 615, 616, 617, 618, 619, 620, 621, 622, 623, 624, 625, 626, 627, 628, 629, 630, 631, 632, 633, 634, 635, 636, 637, 638, 639, 640, 641, 642, 643, 644, 645, 646, 647, 648, 649, 650, 651, 652, 653, 654, 655, 656, 657, 658, 659, 660, 661, 662, 663, 664, 665, 666, 667, 668, 669, 670, 671, 672, 673, 674, 675, 676, 677, 678, 679, 680, 681, 682, 683, 684, 685, 686, 687, 688, 689, 690, 691, 692, 693, 694, 695, 696, 697, 698, 699, or 700 of any of SEQ ID NOS:1-49 and comprising at least, at most, or exactly 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80,

81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193, 194, 195, 196, 197, 198, 199, 200, 201, 202, 203, 204, 205, 206, 207, 208, 209, 210, 211, 212, 213, 214, 215, 216, 217, 218, 219, 220, 221, 222, 223, 224, 225, 226, 227, 228, 229, 230, 231, 232, 233, 234, 235, 236, 237, 238, 239, 240, 241, 242, 243, 244, 245, 246, 247, 248, 249, 250, 251, 252, 253, 254, 255, 256, 257, 258, 259, 260, 261, 262, 263, 264, 265, 266, 267, 268, 269, 270, 271, 272, 273, 274, 275, 276, 277, 278, 279, 280, 281, 282, 283, 284, 285, 286, 287, 288, 289, 290, 291, 292, 293, 294, 295, 296, 297, 298, 299, 300, 301, 302, 303, 304, 305, 306, 307, 308, 309, 310, 311, 312, 313, 314, 315, 316, 317, 318, 319, 320, 321, 322, 323, 324, 325, 326, 327, 328, 329, 330, 331, 332, 333, 334, 335, 336, 337, 338, 339, 340, 341, 342, 343, 344, 345, 346, 347, 348, 349, 350, 351, 352, 353, 354, 355, 356, 357, 358, 359, 360, 361, 362, 363, 364, 365, 366, 367, 368, 369, 370, 371, 372, 373, 374, 375, 376, 377, 378, 379, 380, 381, 382, 383, 384, 385, 386, 387, 388, 389, 390, 391, 392, 393, 394, 395, 396, 397, 398, 399, 400, 401, 402, 403, 404, 405, 406, 407, 408, 409, 410, 411, 412, 413, 414, 415, 416, 417, 418, 419, 420, 421, 422, 423, 424, 425, 426, 427, 428, 429, 430, 431, 432, 433, 434, 435, 436, 437, 438, 439, 440, 441, 442, 443, 444, 445, 446, 447, 448, 449, 450, 451, 452, 453, 454, 455, 456, 457, 458, 459, 460, 461, 462, 463, 464, 465, 466, 467, 468, 469, 470, 471, 472, 473, 474, 475, 476, 477, 478, 479, 480, 481, 482, 483, 484, 485, 486, 487, 488, 489, 490, 491, 492, 493, 494, 495, 496, 497, 498, 499, 500, 501, 502, 503, 504, 505, 506, 507, 508, 509, 510, 511, 512, 513, 514, 515, 516, 517, 518, 519, 520, 521, 522, 523, 524, 525, 526, 527, 528, 529, 530, 531, 532, 533, 534, 535, 536, 537, 538, 539, 540, 541, 542, 543, 544, 545, 546, 547, 548, 549, 550, 551, 552, 553, 554, 555, 556, 557, 558, 559, 560, 561, 562, 563, 564, 565, 566, 567, 568, 569, 570, 571, 572, 573, 574, 575, 576, 577, 578, 579, 580, 581, 582, 583, 584, 585, 586, 587, 588, 589, 590, 591, 592, 593, 594, 595, 596, 597, 598, 599, 600, 601, 602, 603, 604, 605, 606, 607, 608, 609, 610, 611, 612, 613, 614, 615, 616, 617, 618, 619, 620, 621, 622, 623, 624, 625, 626, 627, 628, 629, 630, 631, 632, 633, 634, 635, 636, 637, 638, 639, 640, 641, 642, 643, 644, 645, 646, 647, 648, 649, 650, 651, 652, 653, 654, 655, 656, 657, 658, 659, 660, 661, 662, 663, 664, 665, 666, 667, 668, 669, 670, 671, 672, 673, 674, 675, 676, 677, 678, 679, 680, 681, 682, 683, 684, 685, 686, 687, 688, 689, 690, 691, 692, 693, 694, 695, 696, 697, 698, 699, or 700 (or any derivable range therein) contiguous amino acids or nucleotides of any of SEQ ID NOS:1-49.

**[0050]** The nucleotide as well as the protein, polypeptide, and peptide sequences for various genes have been previously disclosed, and may be found in the recognized computerized databases. Two commonly used databases are the National Center for Biotechnology Information's Genbank and GenPept databases (on the World Wide Web at [ncbi.nlm.nih.gov/](http://ncbi.nlm.nih.gov/)) and The Universal Protein Resource (UniProt; on the World Wide Web at [uniprot.org](http://uniprot.org)). The coding regions for these genes may be amplified and/or expressed using the techniques disclosed herein or as would be known to those of ordinary skill in the art.

**[0051]** It is contemplated that in compositions of the disclosure, there is between about 0.001 mg and about 10 mg of total polypeptide, peptide, and/or protein per ml. The concentration of protein in a composition can be about, at least about or at most about 0.001, 0.010, 0.050, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0, 5.5, 6.0, 6.5, 7.0, 7.5, 8.0, 8.5, 9.0, 9.5, 10.0 mg/ml or more (or any range derivable therein).

**[0052]** In the case of proteins having catalytic activity (e.g., an enzyme), such a protein may be described using Enzyme Classification (EC) nomenclature. EC classifications of various enzymes have been previously disclosed, and may be found in recognized databases, for example, the ENZYME database (Bairoch A. The ENZYME database in 2000. *Nucleic Acids Res.* 2000 Jan 1;28(1):304-5. doi: 10.1093/nar/28.1.304; incorporated herein by reference in its entirety).

## A. Signal Peptides

**[0053]** Aspects of the present disclosure are directed to synthetic signal peptides, and polynucleotides and nucleic acids encoding such signal peptides. Also disclosed are cells comprising such signal peptides, and methods for using cells in production and secretion of a protein (e.g., mammalian protein such as human milk protein). As used herein, "signal peptide" (or "signal peptide sequence") describes any peptide able to, when present at the N-terminal end of a newly synthesized polypeptide, direct the polypeptide across or into a cell membrane of a cell (e.g., the plasma membrane, the endoplasmic reticulum membrane, etc.). In some aspects, a signal peptide of the present disclosure is able to direct a polypeptide into a cell's secretory pathway and subsequent secretion of the polypeptide (described herein as a "secretion signal peptide").

**[0054]** As described herein, aspects of the disclosure relate to synthetic signal peptides comprising:

(a) a pre-region sequence from:

- (i) *P. pastoris* Ost1; or
- (ii) *P. pastoris* Pst1; and

(b) a pro-region sequence from:

- (i) *S. cerevisiae* mating factor  $\alpha$  (MP $\alpha$ ); or  
(ii) *P. pastoris* Epxl.

[0055] Certain signal peptides of the present disclosure are described in Table 1 below.

Table 1 - Signal peptides

Description	Sequence	SEQID NO:
SP1 (pre-Ost1 + pro-MF $\alpha$ )	MKFISILFLLIGSVFGAPVNTTTEDETAQIPAEAVIGYSLEG DFDVAVL PFSNSTNNGLLFINTTASIAAKEEGVSLEKREAE AYVEF	1
SP2 (pre-Pst1 + pro-MF $\alpha$ )	MQFGKVLFAISALAVTALGAPVNTTTEDETAQIPAEAVIGY SDLEGDFDVAVL PFSNSTNNGLLFINTTASIAAKEEGVSLE KREAEAYVEF	2
SP3 (pre-Ost1 + pro-Epx1)	MKFISILFLLIGSVFGAPVAPAEAAANHLHKR	3
SP4 (pre-Pst1 + pro-Epx1)	MQFGKVLFAISALAVTALGAPVAPAEAAANHLHKR	4
SP1 (nucleic acid)	ATGAAATTCATCTCAATTCTGTTCTTTTGATAGGCAGTG TATTTGGTGCTCCAGTCAACACTACAACAGAAGATGAAA CGGCACAAATTCCGGCTGAAGCTGTCATCGGTTACTCAG ATTTAGAAGGGGATTTTCGATGTTGCTGTTTTGCCATTTTC CAACAGCACAAATAACGGGTTATTGTTTATAAATACTAC TATTGCCAGCATTGCTGCTAAAGAAGAAGGGGTATCTCT CGAGAAAAGAGAGGCTGAAGCTTATGTCGAGTTC	41
SP2 (nucleic acid)	ATGCAGTTTGGAAAGGTTCTATTTGCTATTTCTGCCCTGG CTGTCACAGCTCTGGGAGCTCCAGTCAACACTACAACAG AAGATGAAACGGCACAAATTCCGGCTGAAGCTGTCATC GGTTACTCAGATTTAGAAGGGGATTTTCGATGTTGCTGTT TTGCCATTTTCCAACAGCACAAATAACGGGTTATTGTTT ATAAATACTACTATTGCCAGCATTGCTGCTAAAGAAGAA GGGGTATCTCTCGAGAAAAGAGAGGCTGAAGCTTATGTC GAGTTC	42
SP3 (nucleic acid)	ATGAAATTCATCTCAATTCTGTTCTTTTGATAGGCAGTG TATTTGGTGCTCCAGTTGCTCCAGCCGAAGAGGCAGCAA ACCACTTGCACAAGCGT	43

(continued)

Description	Sequence	SEQ ID NO:
SP4 (nucleic acid)	ATGCAGTTTGGAAAGGTTCTATTTGCTATTTCTGCCCTGG CTGTCACAGCTCTGGGAGCTCCAGTTGCTCCAGCCGAAG AGGCAGCAAACCACTTGCACAAGCGT	44

**[0056]** In some aspects, disclosed are polypeptides comprising a signal peptide of the present disclosure. Also disclosed are nucleic acids encoding such polypeptides. Further disclosed are cells expressing polypeptides comprising a signal peptide of the present disclosure.

**[0057]** In some aspects, a polypeptide of the present disclosure comprises SEQ ID NO:1. In some embodiments, a polypeptide of the present disclosure comprises a sequence having at least 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% identity to SEQ ID NO:1. In some aspects, a polypeptide of the present disclosure comprises a sequence having 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10 amino acid substitutions (or more) relative to SEQ ID NO:1.

**[0058]** In some aspects, a polypeptide of the present disclosure comprises SEQ ID NO:2. In some embodiments, a polypeptide of the present disclosure comprises a sequence having at least 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% identity to SEQ ID NO:2. In some aspects, a polypeptide of the present disclosure comprises a sequence having 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10 amino acid substitutions (or more) relative to SEQ ID NO:2.

**[0059]** In some aspects, a polypeptide of the present disclosure comprises SEQ ID NO:3. In some embodiments, a polypeptide of the present disclosure comprises a sequence having at least 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% identity to SEQ ID NO:3. In some aspects, a polypeptide of the present disclosure comprises a sequence having 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10 amino acid substitutions (or more) relative to SEQ ID NO:3.

**[0060]** In some aspects, a polypeptide of the present disclosure comprises SEQ ID NO:4. In some embodiments, a polypeptide of the present disclosure comprises a sequence having at least 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% identity to SEQ ID NO:4. In some aspects, a polypeptide of the present disclosure comprises a sequence having 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10 amino acid substitutions (or more) relative to SEQ ID NO:4.

**[0061]** Any one or more of the signal peptides disclosed herein may be excluded from certain embodiments.

## B. Secretory Proteins

**[0062]** Aspects of the present disclosure include secretory proteins (also "secreted proteins"), as well as compositions comprising secretory proteins, methods of expressing secretory proteins, and methods of use thereof. As used herein, a "secretory protein" describes any protein secreted outside a cell. In certain cases, a secretory protein of the disclosure is a protein present in a human secretion, such as, for example, colostrum, milk, tears, seminal fluid, vaginal fluid, saliva, or other secretion. In some aspects, a secretory protein of the disclosure is a human milk protein. In some aspects, a secretory protein of the disclosure is not a human milk protein.

### 1. Human Milk Proteins

**[0063]** Aspects of the present disclosure include human milk proteins, as well as compositions (e.g., infant formula compositions) comprising human milk proteins, methods of producing human milk proteins, and methods of use thereof. In some aspects, disclosed are cells expressing a human milk protein linked to a signal peptide of the present disclosure (e.g., comprising SEQ ID NOs: 1, 2, 3, or 4). As used herein, a "human milk protein" describes any protein present in human breast milk. A human milk protein includes a protein derived from (e.g., isolated from) human breast milk, as well as any protein produced by other means (e.g., recombinant expression, chemical synthesis, etc.) having an amino acid sequence of a protein present in human breast milk. Various human milk proteins are recognized in the art and contemplated herein. Human milk proteins contemplated herein include, but are not limited to, secretory IgA (sIgA), human serum albumin, xanthine dehydrogenase, lactoferrin, lactoperoxidase, butyrophilin, lactadherin, adiponectin,  $\beta$ -casein,  $\kappa$ -casein, leptin, lysozyme, and  $\alpha$ -lactalbumin. In some embodiments, a human milk protein of the disclosure is a human whey protein. In some embodiments, a human milk protein of the disclosure is a recombinant human milk protein (e.g.,

produced by a non-mammalian cell such as a yeast cell).

**[0064]** Certain aspects of the disclosure are directed to human milk proteins having "human-like" glycans. Human-like glycans (also "human-like glycan structures") describe glycans having structures present in human glycoproteins. Such glycans include, for example, hybrid N-glycans, complex N-glycans, bi-antennary, tri-antennary, and tetra-antennary N-glycans, and glycans comprising sialic acid, galactose, N-acetylgalactosamine, or fucose. Human-like glycans include those having a Man3GlcNAc2 core structure. Accordingly, human milk proteins of the disclosure include those having one or more human-like glycans, for example hybrid N-glycans, complex N-glycans, bi-antennary N-glycans, tri-antennary N-glycans, tetra-antennary N-glycans, and combinations thereof.

**[0065]** Accordingly, in some embodiments, disclosed are recombinant human milk proteins (e.g., recombinant human lactoferrin) comprising one or more human-like glycans. Such recombinant protein include, for example, those produced by engineered mammalian, fungal, yeast, bacterial, or other cells, including engineered cells described elsewhere herein. In certain aspects, such recombinant proteins have a glycan pattern that different from a glycan pattern of a corresponding natural human milk protein. For example, in some embodiments, disclosed is a recombinant human lactoferrin comprising one or more human-like glycans, where the lactoferrin has a glycan pattern that is different from a glycan pattern of any naturally occurring human lactoferrin (e.g., human lactoferrin in human breast milk).

#### a. Lactoferrin

**[0066]** Aspects of the present disclosure are directed to lactoferrin, as well as compositions comprising lactoferrin, including infant formula compositions. In some aspects, disclosed are cells expressing human lactoferrin linked to a signal peptide of the present disclosure (e.g., comprising SEQ ID NOs: 1, 2, 3, or 4). Lactoferrin (also "lactotransferrin") is a whey protein found in exocrine fluids such as breast milk and is encoded by the LTF gene. Without wishing to be bound by theory, lactoferrin is understood to have antimicrobial and anti-inflammatory properties. Certain aspects of the disclosure are directed to human lactoferrin (UniProtKB/Swiss-Prot accession number P02788), including isoforms thereof. The full sequence of human lactoferrin, including signal peptide, is provided as SEQ ID NO:34. The sequence of mature human lactoferrin following cleavage of the signal peptide is provided as SEQ ID NO:9.

Table 2 - Human Lactoferrin sequences

Protein	Sequence	SEQ ID NO
Full length human lactoferrin	MKLVFLVLLFLGALGLCLAGRRRSVQWCAVSQPEATKCFQWQR NMRKVRGPPVSCIKRDSPIQCIQIAIENRADA VTLDGGFIYEAGL APYKLRPVAAEVYGTERQPRTHYYAVAVVKKGGSFQLNELQGL KSCHTGLRRTAGWNVPIGTLRPFLNWTGPPEPIEAAVARFFSASC VPGADKGQFPNLCRLCAGTGENKCAFSSQEPYFSYSGAFKCLRD GAGDVAFIRESTVFEDLSDEAERDEYELLCPDNTRKPVDKFKDC HLARVPSHAVVARSVNGKEDAIWNLLRQAQEKFGKDKSPKFQL FGSPSGQKDLLFKDSAIGFSRVPPRIDSGLYLGSGYFTAIQNLRKS EEEVAARRARVWCAVGEQELRKCNQWSGLSEGSVTCSSASTT EDCIALVLKGEADAMSLDGGYVYTAGKCGLVPVLAENYKSQQS SDPDPNCVDRPVEGYLAVAVVRRSDTSLTWNSVKGKKSCHTAV DRTAGWNIPMGLLFNQGTGCKFDEYFSQSCAPGSDPRSNLCALCI GDEQGENKCVPSNERYYYGYTGAFRCLAENAGDVAFVKDVTVL QNTDGNNNEAWAKDLKLADFALLCLDGKRKPVTEARSCHLAM APNHAVVSRMDKVERLKQVLLHQQAKFGRNGSDCPDKFCLFQS ETKNLLFNDNTECLARLHGKTTYEKYLGPPQYVAGITNLKKCSTS PLLEACEFLRK	34

(continued)

Protein	Sequence	SEQ ID NO
<div>5</div> <div>Mature human lactoferrin</div> <div>10</div> <div>15</div> <div>20</div> <div>25</div>	<div>GRRRSVQWCAVSQPEATKCFQWQRNMRKVRGPPVSCIKRDSPI</div> <div>QCIQAIAENRADAVTLDDGGFIYEAGLAPYKLRPVAAEVYGTERQ</div> <div>PRTHYYAVAVVKKGGSFQLNELQGLKSCHTGLRRTAGWNVPIG</div> <div>TLRPFLNWTGPPEPIEAAVARFFSASCVPAGADKGQFPNLCRLCAG</div> <div>TGENKCAFSSQEPYFSYSGAFKCLRDGAGDVAFIRESTVFEDLSD</div> <div>EAERDEYELLCPDNTRKPVDFKDKCHLARVPSHAVVARSVNGK</div> <div>EDAIWNLLRQAQEFKDKSPKFQLFSGSPGQKDLLFKDSAIGFS</div> <div>RVPPRIDSGLYLGSGYFTAIQNLRKSEEEVAARRARVVWCAVGE</div> <div>QELRKCQWSGLSEGSVTCSSASTTEDCIALVLKGEADAMSLDG</div> <div>GYVYTAGKCGLVPVLAENYKSQQSSDPDPNCVDRPVEGYLAVA</div> <div>VVRRSDTSLTWNSVKGKKSCHTAVDRTAGWNIPMGLLFNQTS</div> <div>CKFDEYFSQSCAPGSDPRSNLALCIGDEQGENKCVPSNERYYY</div> <div>GYTGAFRCLAENAGDVAFVKDVTVLQNTDGNNNEAWAKDLKL</div>	<div>9</div>
<div>30</div> <div>35</div> <div>40</div> <div>45</div> <div>50</div> <div>55</div>	<div>ADFALLCLDGKRKPVTEARSCHLAMAPNHAVVSRMDKVERLK</div> <div>QVLLHQQAKFGRNGSDCPDKFCLFQSETKNLLFNDNTECLARLH</div> <div>GKTTYEKYLGPQYVAGITNLKKCSTSPLEACEFLRK</div>	

(continued)

Protein	Sequence	SEQ ID NO
5 SP1-human lactoferrin  10  15  20  25  30	MKFISILFLLIGSVFGAPVNTTTEDETAQIPAEAVIGYSLEGDFD VAVLPFSNSTNNGLLFINTTIAAIAAKEEGVSLEKREAEAYVEFGR RRSVQWCAVSQPEATKCFQWQRNMRKVRGPPVSCIKRDSPIQCI QAIAENRADAVTLDGGFIYEAGLAPYKLRPVAAEVYGTERTQRT HYYAVAVVKKGGSFQLNELQGLKSCHTGLRRTAGWNVPIGTLR PFLNWTGPPEPIEAAVARFFSASCVPADKGQFPNLCRLCAGTGE NKCAFSSQEPYFSYSGAFKCLRDGAGDVAFIRESTVFEDLSDEAE RDEYELLCPDNTRKPVDFKDKCHLARVPSHAVVARSVNGKEDAI WNLLRQAQEKFGKDKSPKFQLFGSPSGQKDLLFKDSAIGFSRVPP RIDSGLYLGSGYFTAIQNLRKSEEEVAARRARVWCAVGEQELR KCNQWSGLSEGSVTCSSASTTEDCIALLKGEADAMSLDGGYV YTAGKCGLVPVLAENYKSQQSSDPDPNCVDRPVEGYLAVAVVR RSDTSLTWNSVKGKKSCHTAVDRTAGWNIPMGLLFNQTGSCKF DEYFSQSCAPGSDPRSNLCALCIGDEQGENKCVPSNERYYYGYT GAFRCLAENAGDVAFVKDVTVLQNTDGNNEAWAKDLKLADF ALLCLDGKRKPVTEARSCHLAMAPNHAVVSRMDKVERLKQVLL HQQAKFGRNGSDCPDKFCLFQSETKNLLFNDNTECLARLHGKTT YEKYLGPQYVAGITNLKKCSTSPLEACEFLRK	5
35 SP2-human lactoferrin  40  45  50  55	MQFGKVLFAISALAVTALGAPVNTTTEDETAQIPAEAVIGYSLE GDFDVAVLPFSNSTNNGLLFINTTIAAIAAKEEGVSLEKREAEAY VEFGRRSVQWCAVSQPEATKCFQWQRNMRKVRGPPVSCIKRD SPIQCIQAIAENRADAVTLDGGFIYEAGLAPYKLRPVAAEVYGT RQPRTHYYAVAVVKKGGSFQLNELQGLKSCHTGLRRTAGWNV IGTLRPFLNWTGPPEPIEAAVARFFSASCVPADKGQFPNLCRLC AGTGENKCAFSSQEPYFSYSGAFKCLRDGAGDVAFIRESTVFEDL SDEAERDEYELLCPDNTRKPVDFKDKCHLARVPSHAVVARSVN GKEDAIWNLLRQAQEKFGKDKSPKFQLFGSPSGQKDLLFKDSAI GFSRVPPRIDSGLYLGSGYFTAIQNLRKSEEEVAARRARVWCA VGEQELRKCNQWSGLSEGSVTCSSASTTEDCIALLKGEADAMSL DGGYVYTAGKCGLVPVLAENYKSQQSSDPDPNCVDRPVEGYL AVAVVRRSDTSLTWNSVKGKKSCHTAVDRTAGWNIPMGLLFNQT TGSCKFDEYFSQSCAPGSDPRSNLCALCIGDEQGENKCVPSNER YYYGYTGAFRCLAENAGDVAFVKDVTVLQNTDGNNEAWAKDL	6

(continued)

Protein	Sequence	SEQ ID NO
	KLADFALLCLDGKRKPVTEARSCHLAMAPNHAVVSRMDKVERL KQVLLHQQAKFGRNGSDCPDKFCLFQSETKNLLFNDNTECLARL HGKTTYEKYLGPQYVAGITNLKKCSTSPLEACEFLRK	
SP3-human lactoferrin	MKFISILFLLIGSVFGAPVAPAEAAANHLHKRGRRRSVQWCAVSQ PEATKCFQWQRNMRKVVRGPPVSCIQRDSPIQCIQAIENRADAVT LDGGFIYEAGLAPYKLRPVAAEVYGTERRQPRTHYYAVAVVKKG GSFQLNELQGLKSCHTGLRRTAGWNVPIGTLRPFLNWTGPPEPIE AAVARFFSASCVPVGADKGQFPNLCRLCAGTGENKCAFSSQEPYF SYSGAFKCLRDGAGDVAFIRESTVFEDLSDEAERDEYELLCPDNT RKPVDKFKDCHLARVP SHAVVARSVNGKEDAIWNLLRQAQEKF GKDKSPKFQLFGSPSGQKDLLFKDSAIGFSRVPPRIDSGLYLGSGY FTAIQNLRKSEEEVAARRARVWCAVGEQELRKCNCQWSGLSEG SVTCSSASTTEDCIALVLKGEADAMSLDGGYVYTAGKCGLVPVL AENYKSQQSSDPDPNCVDRPVEGYLAVAVVRRSDTSLTWNSVK GKKSCHTAVDRTAGWNIPMGLLFNQGTGCKFDEYFSQSCAPGSD PRSNLCALCIGDEQGENKCVPSNERYYYGYTGAFRCLAENAGDV AFVKDVTVLQNTDGNNNEAWAKDLKLADFALLCLDGKRKPVT EARSCHLAMAPNHAVVSRMDKVERLKQVLLHQQAKFGRNGSD CPDKFCLFQSETKNLLFNDNTECLARLHGKTTYEKYLGPQYVAG ITNLKKCSTSPLEACEFLRK	7



(continued)

Protein	Sequence	SEQ ID NO
5 SP4-human lactoferrin  10  15  20  25	MQFGKVLFAISALAVTALGAPVAPAEAAANHLHKRGRRRSVQW CAVSQPEATKCFQWQRNMRKVRGPPVSCIKRDSPIQCIQAIAENR ADAVTLDGGFIYEAGLAPYKLRPVAAEVYGTERRTHYYAVA VVKKGGSFQLNELQGLKSCHTGLRRTAGWNVPIGTLRPFLNWT GPPEPIEAAVARFFSASCVPAGADKGQFPNLCRLCAGTGENKCAFS SQEPYFSYSGAFKCLRDGAGDVAFIRESTVFEDLSDEAERDEYEL LCPDNTRKPVDKFKDCHLARVP SHAVVARSVNGKEDAIWNLLR QAQEKFGKDKSPKFQLFGSPSGQKDLLFKDSAIGFSRVPPRIDSGL YLGSGYFTAIQNLRKSEEEVAARRARVVWCAVGEQELRKCNQW SGLSEGSVTCSSASTTEDCIALVLKGEADAMSLDGGYVYTAGKC GLVPVLAENYKSQQSSDPDPNCVDRPVEGYLAVAVVRRSDTSLT WNSVKGKKSCHTAVDRTAGWNIPMGLLFNQTGSCKFDEYFSQS CAPGSDPRSNLCALCIGDEQGENKCVPSNERYYYGYTGAFRCLA ENAGDVAFVKDVTVLQNTDGNNNEAWAKDLKLADFALLCLDG KRKPVTEARSCHLAMAPNHAVVSRMDKVERLKQVLLHQQAKF	8
30	GRNGSDCPDKFCLFQSETKNLLFNDNTECLARLHGKTTYEKYLG PQYVAGITNLKKCSTSPLEACEFLRK	

35

40

45

50

55

(continued)

Protein	Sequence	SEQ ID NO
<div>5</div> <div>10</div> <div>15</div> <div>20</div> <div>25</div> <div>30</div> <div>35</div> <div>40</div> <div>45</div> <div>50</div> <div>55</div> <p><i>P. pastoris</i> codon-optimized Human lactoferrin gene</p>	GCCGGAAGAAGAAGAAGTGTTCATGGTGCGCCGTTAGTCAA CCTGAGGCTACAAAGTGTTCATGGCAGAGAAATATGAGA AAGGTTAGAGGTCCACCTGTTTCTTGTATCAAGAGAGATTCTC CAATCCAATGTATTCAAGCTATTGCTGAGAACAGAGCTGATG CTGTTACTTTGGATGGTGGTTTTATCTACGAAGCTGGTTTGGC TCCATATAAACTTAGACCAGTTGCTGCTGAGGTTTACGGTACT GAAAGACAACCTAGAACTCATTACTATGCTGTTGCTGTTGTTA AGAAAGGTGGTTCTTTCCAATTGAACGAATTGCAAGGTTTGA AGTCTTGTCACACTGGTTTGAGAAGAACTGCTGGTTGGAATGT TCCAATTGGTACTTTAAGACCATTCTTAAGTGGACTGGTCCA CCTGAGCCAATTGAAGCTGCTGTTGCTAGATTTTCTCTGCTTC TTGTGTTCCAGGTGCTGATAAGGGTCAATTCCTAATTTGTGT AGATTGTGTGCTGGTACTGGAGAGAACAAATGTGCTTTCTCTT CTCAAGAACCTTACTTTTCTTATTCTGGTGCTTTCAAGTGTTTG AGAGATGGTGCTGGAGATGTTGCTTTTATTAGAGAGTCTACTG TTTTCGAAGATTTGTCTGATGAGGCTGAAAGAGATGAGTATG AATTGTTGTGTCCAGATAACACTAGAAAGCCTGTTGATAAGTT TAAAGATTGTCATTTGGCTAGAGTTCCATCTCACGCTGTTGTT GCTAGATCTGTTAATGGTAAAGAGGATGCTATTTGGAAGTGT TGAGACAAGCTCAAGAAAAGTTCGGTAAAGACAAGTCTCCAA AGTTCCAATTGTTTCGGTTCTCCTTCTGGTCAAAAGGATTTGTT GTTTAAAGATTCTGCTATCGGTTTCTCTAGAGTTCCACCTAGA ATTGATTCTGGTTTGTACTTGGGTTCTGGTTACTTCACTGCTAT CCAAAATTTGAGAAAGTCTGAAGAGGAAGTTGCTGCTAGAAG AGCTAGAGTTGTTTGGTGTGCTGTTGGAGAGCAAGAATTGAG AAAGTGTAACCAATGGTCTGGTTTGTCTGAAGGTTCTGTTACT TGTTCTTCTGCTTCTACTACTGAGGATTGTATTGCTTTGGTTTT GAAAGGTGAAGCTGATGCTATGTCTTTGGATGGTGGTTACGTT TATACTGCTGGTAAATGTGGTTTGGTTCCAGTTTTGGCTGAGA ATTACAAATCTCAACAATCTTCTGATCCAGATCCTAACTGTGT TGATAGACCTGTTGAAGGTTATTTGGCTGTTGCTGTTGTTAGA AGATCTGATACTTCTTTGACTTGGAAGTCTGTAAAGGTAAAA AGTCTTGTCATACTGCTGTTGATAGAACCGCCGGTTGGAATAT TCCAATGGGTTTGTGTTTAACCAAAGTGGTTCTTGTAAGTTT	<div>45</div>

(continued)

Protein	Sequence	SEQ ID NO
5	GATGAGTACTTCTCTCAATCTTGTGCTCCAGGTTCTGATCCTA GATCTAATTTGTGTGCTTTGTGTATTGGAGATGAGCAAGGTGA AAACAAATGTGTTCCCTAATTCTAACGAGAGATACTATGGTTAT	
10	ACTGGTGCTTTTAGATGTTTGGCTGAAAACGCCGGAGATGTTG CTTTCGTTAAGGATGTTACTGTTTTGCAAAACACTGATGGTAA CAATAACGAAGCTTGGGCTAAGGATTTGAAATTGGCTGATTTT	
15	GCTTTGTTGTGTTTGGATGGTAAAAGAAAACCAGTTACTGAGG CTAGATCTTGTCAATTTGGCTATGGCTCCTAACCACGCTGTTGTT TCTAGAATGGATAAGGTTGAAAGATTGAAGCAAGTTTTGTTG	
20	CATCAACAGGCTAAGTTTGGTAGAAATGGTTCTGATTGTCCTG ATAAGTTTTGTTTGTCCAATCTGAGACTAAAACTTGTTGTT CAATGATAACACTGAATGTTTGGCTAGATTGCACGGTAAAAC	
25	TACTTACGAAAAATATTTGGGTCCTCAATACGTTGCTGGTATT ACTAACTTGAAGAAATGCTCCACCAGTCCATTGCTTGAGGCTT GCGAGTTCCTTAGAAAATAA	

30

35

40

45

50

55

(continued)

Protein	Sequence	SEQ ID NO
5 SP1-codon optimized hLF gene  10  15  20  25  30  35	ATGAAATTCATCTCAATTCTGTTCTTTTGATAGGCAGTGTATT TGGTGCTCCAGTCAACACTACAACAGAAGATGAAACGGCACA AATTCCGGCTGAAGCTGTCATCGGTTACTCAGATTTAGAAGGG GATTTTCGATGTTGCTGTTTTGCCATTTTCCAACAGCACAAATA ACGGGTTATTGTTTATAAATACTACTATTGCCAGCATTGCTGC TAAAGAAGAAGGGGTATCTCTCGAGAAAAGAGAGGGCTGAAG CTTATGTCGAGTTCGCCGGAAGAAGAAGAAGTGTTCAATGGT GCGCCGTTAGTCAACCTGAGGCTACAAAGTGTTTTCAATGGCA GAGAAATATGAGAAAGGTTAGAGGTCCACCTGTTTCTTGTATC AAGAGAGATTCTCCAATCCAATGTATTCAAGCTATTGCTGAGA ACAGAGCTGATGCTGTTACTTTGGATGGTGGTTTTATCTACGA AGCTGGTTTTGGCTCCATATAAACTTAGACCAGTTGCTGCTGAG GTTTACGGTACTGAAAGACAACCTAGAACTCATTACTATGCTG TTGCTGTTGTTAAGAAAGGTGGTTCTTTCCAATTGAACGAATT GCAAGGTTTGAAGTCTTGTACACTGGTTTGAGAAGAACTGCT GGTTGGAATGTTCCAATTGGTACTTTAAGACCATTTCTTAACT GGACTGGTCCACCTGAGCCAATTGAAGCTGCTGTTGCTAGATT TTTCTCTGCTTCTTGTGTTCCAGGTGCTGATAAGGGTCAATTC CTAATTTGTGTAGATTGTGTGCTGGTACTGGAGAGAACAAATG TGCTTTCTCTTCTCAAGAACCTTACTTTTCTTATTCTGGTGCTT TCAAGTGTTTGAGAGATGGTGCTGGAGATGTTGCTTTTATTAG	46

40

45

50

55

(continued)

Protein	Sequence	SEQ ID NO
5	AGAGTCTACTGTTTTCGAAGATTTGTCTGATGAGGCTGAAAGA GATGAGTATGAATTGTTGTGTCCAGATAACACTAGAAAGCCT GTTGATAAGTTTAAAGATTGTCATTTGGCTAGAGTTCCATCTC	
10	ACGCTGTTGTTGCTAGATCTGTTAATGGTAAAGAGGATGCTAT TTGGAAC TTGTTGAGACAAGCTCAAGAAAAGTTTCGGTAAAGA CAAGTCTCCAAAGTTCCAATTGTTTCGGTTCTCCTTCTGGTCAA	
15	AAGGATTTGTTGTTTAAAGATTCTGCTATCGGTTTCTCTAGAG TTCCACCTAGAATTGATTCTGGTTTGTACTTGGGTTCTGGTTAC TTCCTGCTATCCAAAATTTGAGAAAGTCTGAAGAGGAAGTT	
20	GCTGCTAGAAGAGCTAGAGTTGTTTGGTGTGCTGTTGGAGAG CAAGAATTGAGAAAGTGTAAACCAATGGTCTGGTTTGTCTGAA GGTTCTGTTACTTGTCTTCTGCTTCTACTACTGAGGATTGTAT	
25	TGCTTTGGTTTTGAAAGGTGAAGCTGATGCTATGTCTTTGGAT GGTGGTTACGTTTATACTGCTGGTAAATGTGGTTTGGTTCCAG TTTTGGCTGAGAATTACAAATCTCAACAATCTTCTGATCCAGA	
30	TCCTAACTGTGTTGATAGACCTGTTGAAGGTTATTTGGCTGTT GCTGTTGTTAGAAGATCTGATACTTCTTTGACTTGGAAGTCTG TTAAAGGTAAAAAGTCTTGTCTACTGCTGTTGATAGAACCGC	
35	CGGTTGGAATATCCAAATGGGTTTGTGTTTAACCAAAGTGGT TCTTGTAAGTTTGATGAGTACTTCTCTCAATCTTGTGCTCCAGG TTCTGATCCTAGATCTAATTTGTGTGCTTTGTGTATTGGAGATG	
40	AGCAAGGTGAAAACAAATGTGTTTCTAATTCTAACGAGAGAT ACTATGGTTATACTGGTGCTTTTAGATGTTTGGCTGAAAACGC CGGAGATGTTGCTTTCGTTAAGGATGTTACTGTTTTGCAAAAC	
45	ACTGATGGTAACAATAACGAAGCTTGGGCTAAGGATTTGAAA TTGGCTGATTTTCGCTTTGTTGTGTTTGGATGGTAAAAGAAAAC CAGTACTGAGGCTAGATCTTGTCAATTTGGCTATGGCTCCTAA	
50	CCACGCTGTTGTTTCTAGAATGGATAAGGTTGAAAGATTGAA GCAAGTTTTGTTGCATCAACAGGCTAAGTTTGGTAGAAATGGT TCTGATTGTCCTGATAAGTTTTGTTTGTTCATCTGAGACTAA	
55	AACTTGTTGTTCAATGATAACACTGAATGTTTGGCTAGATTG CACGGTAAAAC TACTTACGAAAAATATTTGGGTCCTCAATAC GTTGCTGGTATTACTAAGTGAAGAAATGCTCCACCAGTCCAT TGCTTGAGGCTTGCGAGTTCCTTAGAAAATAA	

(continued)

Protein	Sequence	SEQ ID NO
<div>5</div> <div>10</div> <div>15</div> <div>20</div> <div>25</div> <div>30</div> <div>35</div> <div>40</div> <div>45</div> <div>50</div> <div>55</div> SP2-codon optimized hLF gene	ATGCAGTTTGGAAAGGTTCTATTTGCTATTTCTGCCCTGGCTG TCACAGCTCTGGGAGCTCCAGTCAACACTACAACAGAAGATG AAACGGCACAAATTCCGGCTGAAGCTGTCATCGGTTACTCAG ATTTAGAAGGGGATTTTCGATGTTGCTGTTTTGCCATTTTCCAA CAGCACAAATAACGGGTTATTGTTTATAAATACTACTATTGCC AGCATTGCTGCTAAAGAAGAAGGGTATCTCTCGAGAAAAGA GAGGCTGAAGCTTATGTCGAGTTCGCCGGAAGAAGAAGAAGT GTTCAATGGTGCGCCGTTAGTCAACCTGAGGCTACAAAGTGTT TTCAATGGCAGAGAAATATGAGAAAGGTTAGAGGTCCACCTG TTTCTTGTATCAAGAGAGATTCTCCAATCCAATGTATTCAAGC TATTGCTGAGAACAGAGCTGATGCTGTTACTTTGGATGGTGGT TTTATCTACGAAGCTGGTTTGGCTCCATATAAACTTAGACCAG TTGCTGCTGAGGTTTACGGTACTGAAAGACAACCTAGAACTC ATTACTATGCTGTTGCTGTTGTTAAGAAAGGTGGTTCTTTCCA ATTGAACGAATTGCAAGGTTTGAAGTCTTGTCACACTGGTTTG AGAAGAAGTCTGGTTGGAATGTTCCAATTGGTACTTTAAGAC CATTTCTTAACTGGACTGGTCCACCTGAGCCAATTGAAGCTGC TGTTGCTAGATTTTTCTCTGCTTCTTGTGTTCCAGGTGCTGATA AGGGTCAATTTCCCTAATTTGTGTAGATTGTGTGCTGGTACTGG AGAGAACAAATGTGCTTTCTCTTCTCAAGAACCTTACTTTTCT TATTCTGGTGCTTTCAAGTGTTTGAGAGATGGTGCTGGAGATG TTGCTTTTATTAGAGAGTCTACTGTTTTCGAAGATTGTCTGAT GAGGCTGAAAGAGATGAGTATGAATTGTTGTGTCCAGATAAC ACTAGAAAGCCTGTTGATAAGTTTAAAGATTGTCATTTGGCTA GAGTTCCATCTCACGCTGTTGTTGCTAGATCTGTTAATGGTAA AGAGGATGCTATTTGGAAGTTGTTGAGACAAGCTCAAGAAAA GTTTCGGTAAAGACAAGTCTCCAAAGTTCCAATTGTTTCGGTTCT CCTTCTGGTCAAAAGGATTTGTTGTTTAAAGATTCTGCTATCG GTTTCTCTAGAGTTCCACCTAGAATTGATTCTGGTTTGTACTTG GGTTCTGGTTACTTCACTGCTATCCAAAATTTGAGAAAGTCTG AAGAGGAAGTTGCTGCTAGAAGAGCTAGAGTTGTTTGGTGTG CTGTTGGAGAGCAAGAATTGAGAAAGTGTAACCAATGGTCTG GTTTGTCTGAAGGTTCTGTTACTTGTTCTTCTGCTTCTACTACT GAGGATTGTATTGCTTTGGTTTTGAAAGGTGAAGCTGATGCTA TGTCTTTGGATGGTGGTTACGTTTATACTGCTGGTAAATGTGG TTTGGTTCCAGTTTTGGCTGAGAATTACAAATCTCAACAATCT	<div>47</div>

(continued)

Protein	Sequence	SEQ ID NO
5	TCTGATCCAGATCCTAACTGTGTTGATAGACCTGTTGAAGGTT	
	ATTTGGCTGTTGCTGTTGTTAGAAGATCTGATACTTCTTTGACT	
	TGGAACCTCTGTTAAAGGTAAAAAGTCTTGTCATACTGCTGTTG	
10	ATAGAACCGCCGGTTGGAATATTCCAATGGGTTTGTGTTTAA	
	CCAAACTGGTTCTTGTAAGTTTGATGAGTACTTCTCTCAATCTT	
	GTGCTCCAGGTTCTGATCCTAGATCTAATTTGTGTGCTTTGTGT	
15	ATTGGAGATGAGCAAGGTGAAAACAAATGTGTTCCCTAATTCT	
	AACGAGAGATACTATGGTTATACTGGTGCTTTTAGATGTTTGG	
	CTGAAAACGCCGGAGATGTTGCTTTCGTAAAGGATGTTACTGT	
20	TTTGCAAAACACTGATGGTAACAATAACGAAGCTTGGGCTAA	
	GGATTTGAAATTGGCTGATTTCGCTTTGTTGTGTTTGGATGGT	
	AAAAGAAAACCAGTTACTGAGGCTAGATCTTGTCATTTGGCT	
25	ATGGCTCCTAACCACGCTGTTGTTTCTAGAATGGATAAGGTTG	
	AAAGATTGAAGCAAGTTTTGTTGCATCAACAGGCTAAGTTTG	
	GTAGAAATGGTTCTGATTGTCCTGATAAGTTTTGTTTGTCCA	
30	ATCTGAGACTAAAACTTGTTGTTCAATGATAACACTGAATGT	
	TTGGCTAGATTGCACGGTAAACTACTTACGAAAAATATTTGG	
	GTCCTCAATACGTTGCTGGTATTACTAACTTGAAGAAATGCTC	
35	CACCAGTCCATTGCTTGAGGCTTGCGAGTTCCTTAGAAAATAA	

40

45

50

55

(continued)

Protein	Sequence	SEQ ID NO
5 SP3-codon optimized hLF gene  10  15  20  25  30	ATGAAATTCATCTCAATTCTGTTCTTTTGATAGGCAGTGTATT TGGTGCTCCAGTTGCTCCAGCCGAAGAGGCAGCAAACCACTT GCACAAGCGTGCCGGAAGAAGAAGAAGTGTTCAATGGTGCGC CGTTAGTCAACCTGAGGCTACAAAGTGTTTTCAATGGCAGAG AAATATGAGAAAGGTTAGAGGTCCACCTGTTTCTTGTATCAAG AGAGATTCTCCAATCCAATGTATTCAAGCTATTGCTGAGAACA GAGCTGATGCTGTTACTTTGGATGGTGGTTTTATCTACGAAGC TGGTTTGGCTCCATATAAACTTAGACCAGTTGCTGCTGAGGTT TACGGTACTGAAAGACAACCTAGAACTCATTACTATGCTGTTG CTGTTGTTAAGAAAGGTGGTTCCTTTCCAATTGAACGAATTGCA AGGTTTGAAGTCTTGTCACACTGGTTTGAGAAGAAGTCTGGT TGGAATGTTCCAATTGGTACTTTAAGACCATTTCTTAACTGGA CTGGTCCACCTGAGCCAATTGAAGCTGCTGTTGCTAGATTTTT CTCTGCTTCTTGTGTTCCAGGTGCTGATAAGGGTCAATTCCT AATTTGTGTAGATTGTGTGCTGGTACTGGAGAGAACAATGT GCTTTCTCTTCTCAAGAACCTTACTTTTCTTATTCTGGTGCTTT CAAGTGTTTGAGAGATGGTGCTGGAGATGTTGCTTTTATTAGA	48

35

40

45

50

55



(continued)

Protein	Sequence	SEQ ID NO
5	GAGTCTACTGTTTTCGAAGATTTGTCTGATGAGGCTGAAAGAG ATGAGTATGAATTGTTGTGTCCAGATAACACTAGAAAGCCTGT TGATAAGTTTAAAGATTGTCATTTGGCTAGAGTTCCATCTCAC	
10	GCTGTTGTTGCTAGATCTGTTAATGGTAAAGAGGATGCTATTT GGAAC TTGTTGAGACAAGCTCAAGAAAAGTTCGGTAAAGACA AGTCTCCAAAGTTCCAATTGTTTCGGTTCTCCTTCTGGTCAAAA	
15	GGATTTGTTGTTTAAAGATTCTGCTATCGGTTTCTCTAGAGTTC CACCTAGAATTGATTCTGGTTTGTACTTGGGTTCTGGTTACTTC ACTGCTATCCAAAATTTGAGAAAGTCTGAAGAGGAAGTTGCT	
20	GCTAGAAGAGCTAGAGTTGTTTGGTGTGCTGTTGGAGAGCAA GAATTGAGAAAGTGTAACCAATGGTCTGGTTTGTCTGAAGGTT CTGTTACTTGTTCTTCTGCTTCTACTACTGAGGATTGTATTGCT	
25	TTGGTTTTGAAAGGTGAAGCTGATGCTATGTCTTTGGATGGTG GTTACGTTTATACTGCTGGTAAATGTGGTTTGGTTCCAGTTTTG GCTGAGAATTACAAATCTCAACAATCTTCTGATCCAGATCCTA	
30	ACTGTGTTGATAGACCTGTTGAAGGTTATTTGGCTGTTGCTGT TGTTAGAAGATCTGATACTTCTTTGACTTGGAACCTCTGTAAA GGTAAAAAGTCTTGTCACTGCTGTTGATAGAACCGCCGGTT	
35	GGAATATTCCAATGGGTTTGTGTTTAACCAAAGTGGTTCTTG TAAGTTTGATGAGTACTTCTCTCAATCTTGTGCTCCAGGTTCTG ATCCTAGATCTAATTTGTGTGCTTTGTGTATTGGAGATGAGCA	
40	AGGTGAAAACAAATGTGTTCTTAATTCTAACGAGAGATACTA TGGTTATACTGGTGCTTTTAGATGTTTGGCTGAAAACGCCGGA GATGTTGCTTTCGTTAAGGATGTTACTGTTTGCAAAACACTG	
45	ATGGTAACAATAACGAAGCTTGGGCTAAGGATTTGAAATTGG CTGATTTCGCTTTGTTGTGTTTGGATGGTAAAAGAAAACCAAGT TACTGAGGCTAGATCTTGTCAATTTGGCTATGGCTCCTAACCAC	
50	GCTGTTGTTTCTAGAATGGATAAGGTTGAAAGATTGAAGCAA GTTTTGTTGCATCAACAGGCTAAGTTTGGTAGAAATGGTTCTG ATTGTCCTGATAAGTTTTGTTTGTTCCAATCTGAGACTAAAAA	
55	CTTGTTGTTCAATGATAACACTGAATGTTTGGCTAGATTGCAC GGTAAACTACTTACGAAAAATATTTGGGTCCTCAATACGTTG CTGGTATTACTAAGTTGAAGAAATGCTCCACCAGTCCATTGCT TGAGGCTTGCGAGTTCCTTAGAAAATAA	

(continued)

Protein	Sequence	SEQ ID NO
<div>5</div> <div>10</div> <div>15</div> <div>20</div> <div>25</div> <div>30</div> <div>35</div> <div>40</div> <div>45</div> <div>50</div> <div>55</div> SP4-codon optimized hLF gene	ATGCAGTTTGGAAAGGTTCTATTTGCTATTTCTGCCCTGGCTG TCACAGCTCTGGGAGCTCCAGTTGCTCCAGCCGAAGAGGCAG CAAACCACTTGCACAAGCGTGCCGGAAGAAGAAGAAGTGTTT AATGGTGCGCCGTTAGTCAACCTGAGGCTACAAAGTGTGTTTCA ATGGCAGAGAAATATGAGAAAGGTTAGAGGTCCACCTGTTTC TTGTATCAAGAGAGATTCTCCAATCCAATGTATTCAAGCTATT GCTGAGAACAGAGCTGATGCTGTTACTTTGGATGGTGGTTTTA TCTACGAAGCTGGTTTGGCTCCATATAAACTTAGACCAGTTGC TGCTGAGGTTTACGGTACTGAAAGACAACCTAGAACTCATT CTATGCTGTTGCTGTTGTTAAGAAAGGTGGTCTTTCCAATTG AACGAATTGCAAGGTTTGAAGTCTTGTACACTGGTTTGAGAA GAACTGCTGGTTGGAATGTTCCAATTGGTACTTTAAGACCATT TCTTAACTGGACTGGTCCACCTGAGCCAATTGAAGCTGCTGTT GCTAGATTTTTCTCTGCTTCTTGTGTTCCAGGTGCTGATAAGG GTCAATTCCTAATTTGTGTAGATTGTGTGCTGGTACTGGAGA GAACAAATGTGCTTTCTCTTCTCAAGAACCTTACTTTTCTTATT CTGGTGCTTTCAAGTGTTTGAGAGATGGTGCTGGAGATGTTGC TTTTATTAGAGAGTCTACTGTTTTCGAAGATTTGTCTGATGAG GCTGAAAGAGATGAGTATGAATTGTTGTGTCCAGATAACACT AGAAAGCCTGTTGATAAGTTTAAAGATTGTCATTTGGCTAGAG TTCCATCTCACGCTGTTGTTGCTAGATCTGTTAATGGTAAAGA GGATGCTATTTGGAAGTTGTTGAGACAAGCTCAAGAAAAGTT CGGTAAAGACAAGTCTCCAAAGTTCCAATTGTTCCGGTTCTCCT TCTGGTCAAAGGATTTGTTGTTTAAAGATTCTGCTATCGGTT TCTCTAGAGTTCCACCTAGAATTGATTCTGGTTTGTACTTGGG TTCTGGTTACTTCACTGCTATCCAAAATTTGAGAAAGTCTGAA GAGGAAGTTGCTGCTAGAAGAGCTAGAGTTGTTTGGTGTGCT GTTGGAGAGCAAGAATTGAGAAAGTGTAACCAATGGTCTGGT TTGTCTGAAGGTTCTGTTACTTGTCTTCTGCTTCTACTACTGA GGATTGTATTGCTTTGGTTTTGAAAGGTGAAGCTGATGCTATG TCTTTGGATGGTGGTTACGTTTATACTGCTGGTAAATGTGGTT TGGTTCCAGTTTTGGCTGAGAATTACAAATCTCAACAATCTTC TGATCCAGATCCTAACTGTGTTGATAGACCTGTTGAAGGTTAT TTGGCTGTTGCTGTTGTTAGAAGATCTGATACTTCTTTGACTTG GAACTCTGTTAAAGGTAAAAAGTCTTGTCACTGCTGTTGAT AGAACCGCCGGTTGGAATATTCCAATGGGTTTGTGTTTAACC	<div>49</div>

(continued)

Protein	Sequence	SEQ ID NO
5	AAACTGGTTCTTGTAAGTTTGATGAGTACTTCTCTCAATCTTGT	
	GCTCCAGGTTCTGATCCTAGATCTAATTTGTGTGCTTTGTGTAT	
	TGGAGATGAGCAAGGTGAAAACAAATGTGTTTCTAATTCTAA	
10	CGAGAGATACTATGGTTATACTGGTGCTTTTAGATGTTTGGCT	
	GAAAACGCCGGAGATGTTGCTTTTCGTTAAGGATGTTACTGTTT	
	TGCAAAACACTGATGGTAACAATAACGAAGCTTGGGCTAAGG	
15	ATTTGAAATTGGCTGATTTTCGCTTTGTTGTGTTTGGATGGTAA	
	AAGAAAACCAGTTACTGAGGCTAGATCTTGTTCATTTGGCTATG	
	GCTCCTAACCACGCTGTTGTTTCTAGAATGGATAAGGTTGAAA	
20	GATTGAAGCAAGTTTTGTTGCATCAACAGGCTAAGTTTGGTAG	
	AAATGGTTCTGATTGTCCTGATAAGTTTTGTTTGTCCAATCTG	
	AGACTAAAACTTGTGTTCAATGATAACACTGAATGTTTGGC	
25	TAGATTGCACGGTAAACTACTTACGAAAAATATTTGGGTCCT	
	CAATACGTTGCTGGTATTACTAACTTGAAGAAATGCTCCACCA	
	GTCCATTGCTTGAGGCTTGCGAGTTCCTTAGAAAATAA	

**[0067]** In some aspects, a human lactoferrin of the present disclosure is a recombinant human lactoferrin (rhLactoferrin). In some aspects, a recombinant human lactoferrin of the disclosure is obtained from a mammalian, fungal, yeast, bacterial, or other cell. In some aspects, a recombinant human lactoferrin of the disclosure is not obtained from a mammalian cell. In certain aspects, a recombinant human lactoferrin of the disclosure is obtained from a fungal cell. The fungal cell may be, for example, a *Arxula*, *Aspegillus*, *Aurantiochytrium*, *Candida*, *Claviceps*, *Cryptococcus*, *Cunninghamella*, *Geotrichum*, *Hansenula*, *Kluyveromyces*, *Kodamaea*, *Komagataella*, *Leucosporidiella*, *Lipomyces*, *Mortierella*, *Ogataea*, *Pichia*, *Prototheca*, *Rhizopus*, *Rhodospiridium*, *Rhodotorula*, *Saccharomyces*, *Schizosaccharomyces*, *Tremella*, *Trichosporon*, *Wickerhamomyces*, or *Yarrowia* cell. In some aspects, the fungal cell is a yeast cell. In some aspects, the yeast cell is yeast cell is a *Komagataella* cell (e.g., *Komagataella phaffii*, *Komagataella pastoris*, *Komagataella pseudopastoris*). Additional cells suitable for recombinant protein production are recognized in the art and contemplated herein. In some aspects, a recombinant human lactoferrin of the disclosure is obtained from a bacterial cell. In other aspects, a human lactoferrin of the disclosure is isolated from a natural source.

**[0068]** Particular aspects of the present disclosure are directed to human lactoferrin having at least one hybrid or complex N-glycan. In some aspects, the human lactoferrin comprises a glycan comprising one or more of sialic acid, galactose, N-acetylgalactosamine, or fucose. In some aspects, the human lactoferrin comprises a bi-antennary, tri-antennary, or tetra-antennary N-glycan. As disclosed herein, human lactoferrin having one or more hybrid, complex, bi-antennary, tri-antennary, or tetra-antennary N-glycan may be useful in, for example, infant formula or other nutritional compositions or supplements.

#### b. Alpha-lactalbumin ( $\alpha$ -lactalbumin)

**[0069]** Aspects of the present disclosure are directed to alpha-lactalbumin, as well as compositions comprising alpha-lactalbumin, including infant formula compositions. In some aspects, disclosed are cells expressing human alpha-lactalbumin linked to a signal peptide of the present disclosure (e.g., comprising SEQ ID NOs: 1, 2, 3, or 4). Alpha-lactalbumin (also " $\alpha$ -lactalbumin") is a whey protein found in breast milk and is encoded by the LALBA gene. Certain aspects of the disclosure are directed to human  $\alpha$ -lactalbumin (UniProtKB/Swiss-Prot accession number P00709), including isoforms thereof. The full sequence of human  $\alpha$ -lactalbumin, including signal peptide, is provided as SEQ ID NO:36. The sequence of mature human  $\alpha$ -lactalbumin following cleavage of the signal peptide is provided as SEQ ID NO:35.

Table 3 - Human Lactoferrin sequences

Protein	Sequence	SEQ ID NO
Full length human $\alpha$ -lactalbumin	MRFFVPLFLVGILFPAILAKQFTKCELSQLLKDIDGYGGIALPELICTMFHTSGYDTQAIVENNESTEYGLFQISNKLWCKSSQVPQSRNICDISCDKFLDDDDITDDIMCAKKILDIKGIDYWLAHKALCTEKLEQWLCEKL	36
Mature human $\alpha$ -lactalbumin	KQFTKCELSQLLKDIDGYGGIALPELICTMFHTSGYDTQAIVENNESTEYGLFQISNKLWCKSSQVPQSRNICDISCDKFLDDDDITDDIMCAKKILDIKGIDYWLAHKALCTEKLEQWLCEKL	35
SP1-human $\alpha$ -lactalbumin	MKFISILFLLIGSVFGAPVNTTTEDETAQIPAEAVIGYSDLEGDFDVA VLPFSNSTNNGLLFINTTASIAAKEEGVSLEKREAEAYVEFKQFTKCELSQLLKDIDGYGGIALPELICTMFHTSGYDTQAIVENNESTEYGLFQISNKLWCKSSQVPQSRNICDISCDKFLDDDDITDDIMCAKKILDIKGIDYWLAHKALCTEKLEQWLCEKL	37
SP2-human $\alpha$ -lactalbumin	MQFGKVLFAISALAVTALGAPVNTTTEDETAQIPAEAVIGYSDLEGDFDVA VLPFSNSTNNGLLFINTTASIAAKEEGVSLEKREAEAYVEFKQFTKCELSQLLKDIDGYGGIALPELICTMFHTSGYDTQAIVENNESTEYGLFQISNKLWCKSSQVPQSRNICDISCDKFLDDDDITDDIMCAKKILDIKGIDYWLAHKALCTEKLEQWLCEKL	38
SP3-human $\alpha$ -lactalbumin	MKFISILFLLIGSVFGAPVAPAEAAANHLHKKQFTKCELSQLLKDIDGYGGIALPELICTMFHTSGYDTQAIVENNESTEYGLFQISNKLWCKSSQVPQSRNICDISCDKFLDDDDITDDIMCAKKILDIKGIDYWLAHKALCTEKLEQWLCEKL	39
SP4-human $\alpha$ -lactalbumin	MQFGKVLFAISALAVTALGAPVAPAEAAANHLHKKQFTKCELSQLLKDIDGYGGIALPELICTMFHTSGYDTQAIVENNESTEYGLFQISNKLWCKSSQVPQSRNICDISCDKFLDDDDITDDIMCAKKILDIKGIDYWLAHKALCTEKLEQWLCEKL	40

**[0070]** In some aspects, a human  $\alpha$ -lactalbumin of the present disclosure is a recombinant human  $\alpha$ -lactalbumin. In some aspects, a recombinant human  $\alpha$ -lactalbumin of the disclosure is obtained from a mammalian, fungal, yeast, bacterial, or other cell. In some aspects, a recombinant human  $\alpha$ -lactalbumin of the disclosure is not obtained from a mammalian cell. In certain aspects, a recombinant human  $\alpha$ -lactalbumin of the disclosure is obtained from a yeast cell. The yeast cell may be, for example, a *Arxula*, *Aspegillus*, *Aurantiochytrium*, *Candida*, *Claviceps*, *Cryptococcus*, *Cunninghamella*, *Geotrichum*, *Hansenula*, *Kluyveromyces*, *Kodamaea*, *Komagataella*, *Leucosporidiella*, *Lipomyces*, *Mortierella*, *Ogataea*, *Pichia*, *Prototheca*, *Rhizopus*, *Rhodospiridium*, *Rhodotorula*, *Saccharomyces*, *Schizosaccharomyces*, *Tremella*, *Trichosporon*, *Wickerhamomyces*, or *Yarrowia* cell. In some aspects, the yeast cell is yeast cell is a *Komagataella* cell (e.g., *Komagataella phaffii*, *Komagataella pastoris*, *Komagataella pseudopastoris*). Additional yeast cells suit-

able for recombinant protein production are recognized in the art and contemplated herein. In some aspects, a recombinant human  $\alpha$ -lactalbumin of the disclosure is obtained from a bacterial cell. In other aspects, a human  $\alpha$ -lactalbumin of the disclosure is isolated from a natural source.

**[0071]** Particular aspects of the present disclosure are directed to human  $\alpha$ -lactalbumin having at least one hybrid or complex N-glycan. In some aspects, the human  $\alpha$ -lactalbumin comprises a glycan comprising one or more of sialic acid, galactose, N-acetylgalactosamine, or fucose. In some aspects, the human lactoferrin comprises a bi-antennary, tri-antennary, or tetra-antennary N-glycan. As disclosed herein, human  $\alpha$ -lactalbumin having one or more hybrid, complex, bi-antennary, tri-antennary, or tetra-antennary N-glycan may be useful in, for example, infant formula or other nutritional compositions or supplements.

### c. Additional human milk proteins

**[0072]** Additional human milk proteins contemplated in compositions (e.g., infant formula compositions) and methods of the disclosure include, but are not limited to, secretory IgA (sIgA), human serum albumin, xanthine dehydrogenase, lactoperoxidase, butyrophilin, lactadherin, adiponectin,  $\beta$ -casein,  $\kappa$ -casein, leptin, osteopontin, bile salt stimulated lipase (BSSL), and lysozyme. Any one or more of these human milk proteins may be included in compositions (e.g., infant formula) of the present disclosure. Any one or more of these human milk proteins may be excluded in certain embodiments.

### C. N-acetylglucosaminyltransferase

**[0073]** Aspects of the present disclosure relate to an N-acetylglucosaminyltransferase protein. As used herein, an "N-acetylglucosaminyltransferase protein," describes any polypeptide having N-acetylglucosaminyltransferase activity. An N-acetylglucosaminyltransferase describes an enzyme that catalyzes the transfer of a monosaccharide from specific sugar nucleotide donors onto particular hydroxyl position of a monosaccharide in a growing glycan chain in one of two possible anomeric linkages (either  $\alpha$  or  $\beta$ ).

**[0074]** An N-acetylglucosaminyltransferase protein may be an N-acetylglucosaminyltransferase protein from any suitable organism. In some aspects, the N-acetylglucosaminyltransferase protein is a eukaryotic N-acetylglucosaminyltransferase protein. In some aspects, the N-acetylglucosaminyltransferase protein is a mammalian N-acetylglucosaminyltransferase protein.

#### 1. N-acetylglucosaminyltransferase I

**[0075]** In some embodiments, the N-acetylglucosaminyltransferase protein is an N-acetylglucosaminyltransferase I protein (EC 2.4.1.101). The systematic name of this enzyme class is Alpha-1,3-mannosyl-glycoprotein beta-1,2-N-acetylglucosaminyltransferase. Other names include: GnT-I, N-acetylglucosaminyltransferase I, and Uridine diphosphoacetylglucosamine-alpha-1,3-mannosylglycoprotein beta-1,2-N-acetylglucosaminyltransferase. In certain embodiments, an N-acetylglucosaminyltransferase I protein of the present disclosure is *Homo sapiens* GnT-I, however a N-acetylglucosaminyltransferase I protein from any eukaryotic organism may be used as a part of the methods and composition of the disclosure.

#### 2. $\beta$ -1,2-N-acetylglucosaminyltransferase

**[0076]** In some embodiments, the N-acetylglucosaminyltransferase protein is a  $\beta$ -1,2-N-acetylglucosaminyltransferase protein (EC 2.4.1.143). The systematic name of this enzyme class is Alpha-1,6-mannosyl-glycoprotein 2-beta-N-acetylglucosaminyltransferase. Other names include: GnT-II, N-acetylglucosaminyltransferase II, and Uridine diphosphoacetylglucosamine-alpha-1,6-mannosylglycoprotein beta-1,2-N-acetylglucosaminyltransferase. In certain embodiments, a  $\beta$ -1,2-N-acetylglucosaminyltransferase protein of the present disclosure is *Rattus norvegicus* GnT-II, however a  $\beta$ -1,2-N-acetylglucosaminyltransferase protein from any eukaryotic organism may be used as a part of the methods and composition of the disclosure.

### D. Alpha-1,3/6-Mannosidase ( $\alpha$ -1,3/6-Mannosidase)

**[0077]** Aspects of the present disclosure relate to an  $\alpha$ -1,3/6-Mannosidase protein (EC 3.2.114). As used herein, an " $\alpha$ -1,3/6-Mannosidase protein" (or "alpha-1,3/6-Mannosidase protein") describes any polypeptide having  $\alpha$ -1,3/6-Mannosidase activity. An  $\alpha$ -1,3/6-Mannosidase describes an enzyme that catalyzes removal of two mannosyl residues from N-glycans. The systematic name of this enzyme class is Mannosyl-oligosaccharide 1,3-1,6-alpha-mannosidase. Other names include: Man-II and Mannosidase II. An  $\alpha$ -1,3/6-Mannosidase protein may be from any suitable organism. In some embodiments, the  $\alpha$ -1,3/6-Mannosidase protein is a eukaryotic  $\alpha$ -1,3/6-Mannosidase protein. In certain embodi-

ments, the  $\alpha$ -1,3/6-Mannosidase protein is *Drosophila melanogaster* Man-II, however a  $\alpha$ -1,3/6-Mannosidase protein from any eukaryotic organism may be used as a part of the methods and composition of the disclosure.

#### E. Alpha-1,2-mannosidase ( $\alpha$ -1,2-mannosidase)

[0078] Aspects of the present disclosure relate to a  $\alpha$ -1,2-mannosidase protein (EC 3.2.1.130). As used herein, a " $\alpha$ -1,2-mannosidase protein" (or "alpha-1,2-mannosidase protein") describes any polypeptide having  $\alpha$ -1,2-mannosidase activity. The systematic name of this enzyme class is Glycoprotein endo-alpha-1,2-mannosidase. Other names include: Endo-alpha-D-mannosidase and Man-I. In some embodiments, the  $\alpha$ -1,2-mannosidase protein is a fungal Man-I. In certain embodiments, the Man-I is a *Trichoderma reesei* Man-I.

#### F. Beta-1,4- galactosyltransferase ( $\beta$ -1,4-galactosyltransferase)

[0079] Aspects of the present disclosure relate to a  $\beta$ -1,4-galactosyltransferase protein (EC 2.4.1.38). As used herein, a " $\beta$ -1,4-galactosyltransferase protein" (or "beta-1,4-galactosyltransferase protein") describes any polypeptide having  $\beta$ -1,4-galactosyltransferase activity. The systematic name of this enzyme class is Beta-N-acetylglucosaminylglycopeptide beta-1,4-galactosyltransferase. Other names include: Glycoprotein 4-beta-galactosyltransferase, UDP-galactose--glycoprotein galactosyltransferase, and GalT. In some embodiments, the  $\beta$ -1,4-galactosyltransferase protein is a mammalian GalT. In certain embodiments, the GalT is a *Homo Sapiens* GalT.

#### G. Glycosylated Proteins

[0080] Aspects of the present disclosure are directed to methods and compositions for production of glycosylated proteins (also "glycoproteins") having patterns of glycosylation similar to those of glycoproteins produced by human cells. In some embodiments, glycoproteins of the disclosure are N-linked glycoproteins. N-linked glycoproteins contain an N-acetylglucosamine residue linked to the amide nitrogen of an asparagine residue in the protein. The predominant sugars found on glycoproteins are glucose, galactose, mannose, fucose, N-acetylgalactosamine (GalNAc), N-acetylglucosamine (GlcNAc), and sialic acid, e.g., N-acetyl-neuraminic acid (NANA). The processing of the sugar groups occurs co-translationally in the lumen of the ER and continues in the Golgi apparatus for N-linked glycoproteins.

#### H. Protein targeting

[0081] Certain aspects of the present disclosure include cells expressing one or more proteins from a nucleic acid molecule, where the protein is targeted to a desired subcellular location (e.g., an organelle such as the Golgi Apparatus). In some cases, a protein is targeted to a subcellular location by forming a fusion protein comprising a portion of the protein (e.g., a catalytic domain of an enzyme) and a cellular targeting signal peptide, e.g., a heterologous signal peptide (e.g., a signal peptide comprising SEQ ID NO:1, 2, 3, or 4) which is not normally ligated to or associated with the portion of the protein. The fusion protein may be encoded by a polynucleotide encoding a cellular targeting signal peptide ligated in the same translational reading frame ("in-frame") to a nucleic acid fragment encoding a protein (e.g., enzyme), or catalytically active fragment thereof.

[0082] The targeting signal peptide component of the fusion construct or protein may be derived from membrane-bound proteins of the ER or Golgi, retrieval signals, Type II membrane proteins, Type I membrane proteins, membrane spanning nucleotide sugar transporters, mannosidases, sialyltransferases, glucosidases, mannosyltransferases and phosphomannosyltransferases. In some aspects, the targeting signal peptide is a Golgi Apparatus localization tag. Example Golgi Apparatus localization tags include, but are not limited to, a transmembrane domain from *Saccharomyces cerevisiae* Kre2p, *Saccharomyces cerevisiae* Mnn2p, *Saccharomyces cerevisiae* Mnn9, *Komagatella phaffii* Bmt2, *Komagatella phaffii* Bmt3, or *Komagatella phaffii* Ktr2.

#### III. Sequences

[0083] Certain example polypeptide and nucleic sequences contemplated herein are shown below in Table 4.

Table 4

Description	Sequence	SEQ ID NO:
SP1	MKFISILFLLIGSVFGAPVNTTTEDETAQIPAEAVIGYSDLEGDFDVAVLPF SNSTNNGLLFINTTIIASIAAKEEGVSLEKREAEAYVEF	1
SP2	MQFGKVLFAISALAVTALGAPVNTTTEDETAQIPAEAVIGYSDLEGDFDV AVLPFSNSTNNGLLFINTTIIASIAAKEEGVSLEKREAEAYVEF	2
SP3	MKFISILFLLIGS V FGAPV APAEEAANHLHKR	3
SP4	MQFGKVLFAISALAVTALGAPVAPAEAAANHLHKR	4
SP1-hLF	<u>MKFISILFLLIGSVFGAPVNTTTEDETAQIPAEAVIGYSDLEGDFDVAVLPF</u> <u>SNSTNNGLLFINTTIIASIAAKEEGVSLEKREAEAYVEF</u> GRRRSVQWCAVS QPEATKCFQWQRNMRKVRGPPVSCIQRDSPIQCIQAIENRADA VTLDG GFIYEAGLAPYKLRPVAAEVYGTERTPRTHYYAVAVVKKGGSFQNL QGLKSCHTGLRRTAGWNVPIGTLRPFLNWTGPPEPIEA AVARFFSASCVP GADKGQFPNLCRLCAGTGENKCAFSSQEPYFSYSGAFKCLRDGAGD VAF IRESTVFEDLSDEAERDEYELLCPDNTRKPVDFKDKCHLARVP SHAVVA RSVNGKEDAIWNLLRQAQEKFGKDKSPKFQLFGPSGQKDLLFKD SAIG FSRVPPRIDSGLYLGSGYFTAIQNLRKSEEEVAARRARVVWCAV GEQEL RKCQNSGLSEGSVTCSSASTTEDCIALVLKGEADAMSLDGGYVY TAG KCGLPVLAENYKSQQSSDPDPNCVDRPVEGYLAVAVVRRSDT SLTWN SVKGKKSCHTAVDRTAGWNIPMGLLFNQGTGCKFDEYFSQSCAP GSDPR SNLCALCIGDEQGENKCVPSNERYYYGYTGAFRCLAENAGDVAFV KDV TVLQNTDGNNEAWAKDLKLADFALLCLDGKRKPVTEARSCHLA MAP NHAVVSRMDKVERLKQVLLHQQAKFGRNGSDCPDKFCLFQSETK NLLF NDNTECLARLHGKTTYEYKYLGPQYVAGITNLKKCSTSPLEACE FLRK	5
SP2-hLF	<u>MQFGKVLFAISALAVTALGAPVNTTTEDETAQIPAEAVIGYSDLEGDFDV</u> <u>AVLPFSNSTNNGLLFINTTIIASIAAKEEGVSLEKREAEAYVEF</u> GRRRSVQ WCAVSQPEATKCFQWQRNMRKVRGPPVSCIQRDSPIQCIQAIENRADA VTLDGGFIYEAGLAPYKLRPVAAEVYGTERTPRTHYYAVAVVKKGGSF QLNELQGLKSCHTGLRRTAGWNVPIGTLRPFLNWTGPPEPIEA AVARFFS ASCVP GADKGQFPNLCRLCAGTGENKCAFSSQEPYFSYSGAFKCLRDGA GDVAFIRESTVFEDLSDEAERDEYELLCPDNTRKPVDFKDKCHLARVP SH AVVARSVNGKEDAIWNLLRQAQEKFGKDKSPKFQLFGPSGQKDLLFK DSAIGFSRVPPRIDSGLYLGSGYFTAIQNLRKSEEEVAARRARVVWCAV G EQELRKCQNSGLSEGSVTCSSASTTEDCIALVLKGEADAMSLDGGYVY	6
	TAGKCGLPVLAENYKSQQSSDPDPNCVDRPVEGYLAVAVVRRSDTSLT WNSVKGKKSCHTAVDRTAGWNIPMGLLFNQGTGCKFDEYFSQSCAPGS DPRSNLCALCIGDEQGENKCVPSNERYYYGYTGAFRCLAENAGDVAFV KDVTVLQNTDGNNEAWAKDLKLADFALLCLDGKRKPVTEARSCHLA MAPNHAVVSRMDKVERLKQVLLHQQAKFGRNGSDCPDKFCLFQSETKN LLFNDNTECLARLHGKTTYEYKYLGPQYVAGITNLKKCSTSPLEACEFLR K	

(continued)

Description	Sequence	SEQ ID NO:
SP3-hLF	<u>MKFISILFLLIGSVFGAPVAPAEAAANHLHKGRRRSVQWCAVSQPEATK</u> CFQWQRNMRKVRGPPVSCIKRDSPIQCIQAIENRADAVTLDGGFIYEAG LAPYKLRPVAAEVYGTERTPRTHYYAVAVVKGGSFQNLQGLKSCH TGLRRTAGWNVPIGTLRPFLNWTGPPEPIEAAVARFFSASCVPADKGGQF PNLRLCAGTGENKCAFSSQEPYFSYSGAFKCLRDGAGDVAFIRESTVFE DLSDEAERDEYELLCPDNTRKPVDFKDKCHLARVPSHAVVARSVNGKE DAIWNLLRQAQEKFGKDKSPKFQFLGSPSGQKDLLFKDSAIGFSRVPPRI DSGLYLGSGYFTAIQNLRKSEEEVAARRARVVWCAVGEQELRKCNQWS GLSEGSVTCSSASTTEDICIALVLKGEADAMSLDGGYVYTAGKCGLPVL AENYKSQQSSDPDPNCVDRPVEGYLAVAVVRRSDTSLTWNSVKGKKS HTAVDRTAGWNIPMGLLFNQTGCKFDEYFSQSCAPGSDPRSNLCALCI GDEQGENKCVPSNERYGYTGAFRCLAENAGDVAFVKDVTVLQNTD GNNNEAWAKDLKLADFALLCLDGKRKPVTEARSCHLAMAPNHAVVSR MDKVERLKQVLLHQQAKFGRNGSDCPDKFCLFQSETKNLLFNDNTECL ARLHGKTTYEKYLGOPYVAGITNLKKCSTSPLEACEFLRK	7
SP4-hLF	<u>MOFGKVLFAISALAVTALGAPVAPAEAAANHLHKGRRRSVQWCAVSQ</u> PEATKCFQWQRNMRKVRGPPVSCIKRDSPIQCIQAIENRADAVTLDGGFI IYEAGLAPYKLRPVAAEVYGTERTPRTHYYAVAVVKGGSFQNLQGLQ LKSGHTGLRRTAGWNVPIGTLRPFLNWTGPPEPIEAAVARFFSASCVP DKGQFPNLRLCAGTGENKCAFSSQEPYFSYSGAFKCLRDGAGDVAFIR ESTVFEDLSDEAERDEYELLCPDNTRKPVDFKDKCHLARVPSHAVVARS VNGKEDAIWNLLRQAQEKFGKDKSPKFQFLGSPSGQKDLLFKDSAIGFS RVPPRIDSGLYLGSGYFTAIQNLRKSEEEVAARRARVVWCAVGEQELRK CNQWSGLSEGSVTCSSASTTEDICIALVLKGEADAMSLDGGYVYTAGKC GLVPVLAENYKSQQSSDPDPNCVDRPVEGYLAVAVVRRSDTSLTWNSV KGKKSCHTAVDRTAGWNIPMGLLFNQTGCKFDEYFSQSCAPGSDPRSN LCALCIGDEQGENKCVPSNERYGYTGAFRCLAENAGDVAFVKDVTVLQ NTDGNNEAWAKDLKLADFALLCLDGKRKPVTEARSCHLAMAPNH AVVSRMDKVERLKQVLLHQQAKFGRNGSDCPDKFCLFQSETKNLLFND NTECLARLHGKTTYEKYLGOPYVAGITNLKKCSTSPLEACEFLRK	8
hLF	GRRRSVQWCAVSQPEATKCFQWQRNMRKVRGPPVSCIKRDSPIQCIQAI AENRADAVTLDGGFIYEAGLAPYKLRPVAAEVYGTERTPRTHYYAVAV VKGGSFQNLQGLKSGHTGLRRTAGWNVPIGTLRPFLNWTGPPEPIEAA VARFFSASCVPADKGGQFPNLRLCAGTGENKCAFSSQEPYFSYSGAF KCLRDGAGDVAFIRESTVFEDLSDEAERDEYELLCPDNTRKPVDFKDC HLARVPSHAVVARSVNGKEDAIWNLLRQAQEKFGKDKSPKFQFLGSPSG QKDLLFKDSAIGFSRVPPRIDSGLYLGSGYFTAIQNLRKSEEEVAARRAR VVWCAVGEQELRKCNQWSGLSEGSVTCSSASTTEDICIALVLKGEADAM SLDGGYVYTAGKCGLPVLAENYKSQQSSDPDPNCVDRPVEGYLAVAV VRRSDTSLTWNSVKGKKSCHTAVDRTAGWNIPMGLLFNQTGCKFDEY FSQSCAPGSDPRSNLCALCIGDEQGENKCVPSNERYGYTGAFRCLAE NAGDVAFVKDVTVLQNTDGNNEAWAKDLKLADFALLCLDGKRKPVTE EARSCHLAMAPNHAVVSRMDKVERLKQVLLHQQAKFGRNGSDCPDKF CLFQSETKNLLFNDNTECLARLHGKTTYEKYLGOPYVAGITNLKKCSTSP LLEACEFLRK	9
S. cerevisiae pro-MF $\alpha$ (1)	MRFPSIFTAVLFAASSALA	10
S. cerevisiae pro-MF $\alpha$ (2)	APVNTTTEDETAQIPAEAVIGYSLEGDGFDVAVLPFSNSTNGLLFINTTI ASIAAKEEGVSLEKREAEAYVEF	11
P. pastoris Ost1	MKFISILFLLIGSVFG	12



EP 4 424 698 A2

(continued)

Description	Sequence	SEQ ID NO:
P. pastoris Epx1 pro region	APVAPAEAAANHLHKR	13
Pichia pastoris Pst1	MQFGKVLFAISALAVTALG	14
gBLOCK1	ATGCAGTTTGGAAAGGTTCTATTTGCTATTTCTGCCCTGGCTGTCACAGCTCTGGGAGCTCCAGTTGCTCCAGCCGAAGAGGCAGCAAACCACTTGCACAAGCGTATGAGGCAGGTTTGGTTCTCTTGGATTGTGGGATTGTT CCTATGTTTTTTCAACGTGTCTTCTGCTAAACGATGAAATTCATCTCAATTCTGTTCCTTTTGATAGGCAGTGTATTTGGTATGAAATTCATCTCAATTCTGTTCCTTTTGATAGGCAGTGTATTTGGTGCTCCAGTTGCTCCAGCCGAAGAGGCAGCAAACCACTTGCACAAGCGT	15
PMR1	CGAAGGATCCAAACGATGAAATTCATCTCAATTC	16
PMR2	GTAGTGTTGACTGGAGCACCAAATACAC	17
PMR3	GGTGCTCCAGTCAACACTACAACAGAAG	18
PMR4	TTCATCGTTTGGATCCTTCGAATAATTAGTTG	19
PMR5	CGAAGGATCCAAACGATGCAGTTTGGAAAGG	20
PMR6	TGACTGGAGCTCCCAGAGCTGTGACAGC	21
PMR7	AGCTCTGGGAGCTCCAGTCAACACTACAAC	22
PMR8	TGCATCGTTTGGATCCTTCGAATAATTAGTTGTTTTTTG	23
PMR9	GATCCAAACGATGAAATTCATCTCAATTCTGTTCCCTTTTG	24
PMR10	TTCTTCCGGGCACGCTTGTGCAAGTGGTTTG	25
PMR11	GCACAAGCGTGCCGGAAGAAGAAGAAGTG	26
PMR12	TGAATTTTCATCGTTTGGATCCTTCGAATAATTAG	27
PMR13	GATCCAAACGATGCAGTTTGGAAAGGTTCTATTTG	28
PMR14	TTCTTCCGGGCACGCTTGTGCAAGTGGTTTG	29
PMR15	GCACAAGCGTGCCGGAAGAAGAAGAAGTG	30
PMR16	CAAACGTCATCGTTTGGATCCTTCGAATAATTAG	31
PMR17	GATCTAACATCCAAAGACGAAA	32
PMR18	TTGAGATAAATTTACGTTTAA	33

(continued)

Description	Sequence	SEQ ID NO:
Full-length human lactoferrin (hLF)	MKLVFLVLLFLGALGLCLAGRRRSVQWCAVSQPEATKCFQWQRNMRK VRGPPVSCIKRDSPIQCIQAIENRADAVTLDGGFIYEAGLAPYKLRPVAA EVYGTERRQPRTHYYAVAVVKKGGSFQLNELQGLKSCHTGLRRTAGWN VPIGTLRPFLNWTGPPEPIEAAVARFFSASCVPGADKGQFPNLCRLCAGT GENKCAFSSQEPYFSYSGAFKCLRDGAGDVAFIRESTVFEDLSDEAERDE YELCPDNTRKPVDFKDKCHLARVP SHAVVARSVNGKEDAIWNLLRQA QEKFGKDKSPKFQFLFGSPSGQKDLLFKDSAIGFSRVPPRIDSGLYLGSGYF TAIQNLRKSEEEVAARRARVWCAVGEQELRKCNQWSGLSEGSVTCSS ASTTEDCIALVLKGEADAMSLDGGYVYTAGKCGLVPVLAENYKSQQSS DPDPNCVDRPVEGYLA VAVVRRSDTSLTWN SVKGKKSCHTAVDRTAG WNIPMGLLNFNTGSCKFDEYFSQSCAPGSDPRSNLCALCIGDEQGENKC VPNSNERYYGYTGAFRCLAENAGDVA FVKDVTVLQNTDGNNEAWAK DLKLADFALLCLDGKRKPVTEARSCHLAMAPNHAVVSRMDKVERLKQ VLLHQQAKFGRNGSDCPDKFCLFQSETKNLLFNDNTECLARLHGKTTYE KYLGPQYVAGITNLKKCSTSPLEACEFLRK	34
Human alpha-lactalbumin (hALA)	KQFTKCELSQLLKDIDGYGGIALPELICTMFHTSGYDTQAIVENNESTEY GLFQISNKLWCKSSQVPQSRNICDISCDKFLDDITDDIMCAKKILDIKGI DYWLAHKALCTEKLQWLCEKL	35
Full-length hALA	MRFFVPLFLVGILFPAILAKQFTKCELSQLLKDIDGYGGIALPELICTMFHT SGYDTQAIVENNESTEYGLFQISNKLWCKSSQVPQSRNICDISCDKFLDD DITDDIMCAKKILDIKIDYWLAHKALCTEKLQWLCEKL	36
SP1-hALA	MKFISILFLLIGSVFGAPVNTTTEDETAQIPAEAVIGYSDLEGDFDVAVLPF SNSTNNGLLFINTTIASIAAKEEGVSLEKREAEAYVEFKQFTKCELSQLLK DIDGYGGIALPELICTMFHTSGYDTQAIVENNESTEYGLFQISNKLWCKSS QVPQSRNICDISCDKFLDDITDDIMCAKKILDIKIDYWLAHKALCTEKL LEQWLCEKL	37
SP2-hALA	MQFGKVLFAISALAVTALGAPVNTTTEDETAQIPAEAVIGYSDLEGDFDV AVLPFSNSTNNGLLFINTTIASIAAKEEGVSLEKREAEAYVEFKQFTKCEL SQLLKDIDGYGGIALPELICTMFHTSGYDTQAIVENNESTEYGLFQISNKL WCKSSQVPQSRNICDISCDKFLDDITDDIMCAKKILDIKIDYWLAHKA LCTEKLQWLCEKL	38
SP3-hALA	MKFISILFLLIGSVFGAPVAPAEAAANHLHKKRQFTKCELSQLLKDIDGYG GIALPELICTMFHTSGYDTQAIVENNESTEYGLFQISNKLWCKSSQVPQSR NICDISCDKFLDDITDDIMCAKKILDIKIDYWLAHKALCTEKLQWLC EKL	39
SP4-hALA	MQFGKVLFAISALAVTALGAPVAPAEAAANHLHKKRQFTKCELSQLLKD IDGYGGIALPELICTMFHTSGYDTQAIVENNESTEYGLFQISNKLWCKSSQ VPQSRNICDISCDKFLDDITDDIMCAKKILDIKIDYWLAHKALCTEKL QWLCEKL	40
SP1 (nucleic acid)	ATGAAATTCATCTCAATTCTGTTCTTTTGATAGGCAGTGTATTTGGT GCTCCAGTCAACACTACAACAGAAGATGAAACGGCACAAATTCGGC TGAAGCTGTCATCGGTTACTCAGATTTAGAAGGGGATTTCGATGTTGC TGTTTTGCCATTTTCCAACAGCACAAATAACGGGTATTGTTTATAAA TACTACTATTGCCAGCATTGCTGCTAAAGAAGAAGGGGTATCTCTCG AGAAAAGAGAGGCTGAAGCTTATGTCGAGTTC	41

(continued)

Description	Sequence	SEQ ID NO:
SP2 (nucleic acid)	ATGCAGTTTGGAAAGGTTCTATTTGCTATTTCTGCCCTGGCTGTCACA GCTCTGGGAGCTCCAGTCAACACTACAACAGAAGATGAAACGGCACA AATTCGGGCTGAAGCTGTCATCGGTTACTCAGATTTAGAAGGGGATTT CGATGTTGCTGTTTTGCCATTTTCCAACAGCACAAATAACGGGTTATT GTTTATAAATACTACTATTGCCAGCATTTGCTGCTAAAGAAGAAGGGG TATCTCTCGAGAAAAGAGAGGCTGAAGCTTATGTCGAGTTC	42
SP3 (nucleic acid)	ATGAAATTCATCTCAATTCTGTTCTTTTGATAGGCAGTGTATTTGGT GCTCCAGTTGCTCCAGCCGAAGAGGCAGCAAACCACTTGCACAAGCG T	43
SP4 (nucleic acid)	ATGCAGTTTGGAAAGGTTCTATTTGCTATTTCTGCCCTGGCTGTCACA GCTCTGGGAGCTCCAGTTGCTCCAGCCGAAGAGGCAGCAAACCACTT GCACAAGCGT	44
Codon-optimized hLF	GCCGGAAGAAGAAGAAGTGTTC AATGGTGCGCCGTTAGTCAACCTGA GGCTACAAAGTGTTC AATGGCAGAGAAATATGAGAAAGGTTAGAG GTCCACCTGTTTCTTGTATCAAGAGAGATTCTCCAATCCAATGTATTC AAGCTATTGCTGAGAACAGAGCTGATGCTGTTACTTTGGATGGTGGTT TTATCTACGAAGCTGGTTTGGCTCCATATAAACTTAGACCAGTTGCTG CTGAGGTTTACGGTACTGAAAGACAACCTAGAACTCATTACTATGCT GTTGCTGTTGTTAAGAAAGGTGGTTCTTTCCAATTGAACGAATTGCAA GGTTTGAAGTCTTGTCCACTGGTTTGAGAAGAACTGCTGGTTGGAAT GTTCCAATTGGTACTTTAAGACCATTCTTAACTGGACTGGTCCACCT GAGCCAATTGAAGCTGCTGTTGCTAGATTTTCTCTGCTTCTTGTGTTT CAGGTGCTGATAAGGGTCAATTCCTAATTTGTGTAGATTGTGTGCTG GTACTGGAGAGAAACAAATGTGCTTTCTCTTCTCAAGAACCTTACTTTT CTTATTCTGGTGCTTTCAAGTGTTCAGAGATGGTGCTGGAGATGTTG CTTTTATTAGAGAGTCTACTGTTTTCGAAGATTTGTCTGATGAGGCTG AAAGAGATGAGTATGAATTGTTGTGTCCAGATAACACTAGAAAGCCT GTTGATAAGTTTAAAGATTGTCATTTGGCTAGAGTTCCATCTCACGCT GTTGTTGCTAGATCTGTTAATGGTAAAGAGGATGCTATTTGGAACCTG TTGAGACAAGCTCAAGAAAAGTTCGGTAAAGACAAGTCTCCAAAGTT CCAATTGTTTCGGTTCTCCTTCTGGTCAAAGGATTTGTTGTTTAAAGA TTCTGCTATCGGTTTCTCTAGAGTTCCACCTAGAATTGATTCTGGTTTG TACTTGGGTTCTGGTTACTTCACTGCTATCCAAAATTTGAGAAAGTCT GAAGAGGAAGTTGCTGCTAGAAGAGCTAGAGTTGTTTGGTGTGCTGT TGGAGAGCAAGAATTGAGAAAGTGTAACCAATGGTCTGGTTTGTCTG AAGGTTCTGTTACTTGTTCCTGCTTCTACTACTGAGGATTGTATTGC	45

(continued)

Description	Sequence	SEQ ID NO:
	TTTGGTTTTGAAAGGTGAAGCTGATGCTATGTCTTTGGATGGTGGTTA CGTTTATACTGCTGGTAAATGTGGTTTTGGTTCCAGTTTTGGCTGAGAA TTACAAATCTCAACAATCTTCTGATCCAGATCCTAACTGTGTTGATAG ACCTGTTGAAGGTTATTTGGCTGTTGCTGTTGTTAGAAGATCTGATAC TTCTTTGACTTGGAACCTCTGTTAAAGGTAAAAAGTCTTGTCATACTGC TGTTGATAGAACCGCCGGTTGGAATATTCCAATGGGTTTGTGTTTAA CCAAACCTGGTTCTTGTAAGTTTGATGAGTACTTCTCTCAATCTTGTCCT CCAGGTTCTGATCCTAGATCTAATTTGTGTGCTTTGTGTATTGGAGAT GAGCAAGGTGAAAACAAATGTGTTTCTAATTCTAACGAGAGATACTA TGGTTATACTGGTGCCTTTTAGATGTTTGGCTGAAAACGCCGGAGATGT TGCTTTCGTTAAGGATGTTACTGTTTTGCAAAACACTGATGGTAACAA TAACGAAGCTTGGGCTAAGGATTTGAAATTGGCTGATTCGCTTTGTT GTGTTTGGATGGTAAAAGAAAACAGTTACTGAGGCTAGATCTTGTC ATTTGGCTATGGCTCCTAACACGCTGTTGTTTCTAGAATGGATAAGG TTGAAAGATTGAAGCAAGTTTTGTTGCATCAACAGGCTAAGTTTGTA GAAATGGTTCTGATTGTCCTGATAAGTTTTGTTTGGTCCAATCTGAGA CTAAAAACTTGTGTTCAATGATAACACTGAATGTTTGGCTAGATTGC ACGGTAAAACCTTACGAAAAATATTTGGGTCCTCAATACGTTGCTG GTATTACTAACTTGAAGAAATGCTCCACCAGTCCATTGCTTGAGGCTT GCGAGTTCCTTAGAAAATAA	

(continued)

Description	Sequence	SEQ ID NO:
SP1-codon optimized hLF gene	ATGAAATTCATCTCAATTCTGTTTCCTTTTGATAGGCAGTGTATT TGGTGCTCCAGTCAACACTACAACAGAAGATGAAACGGCACA AATTCCGGCTGAAGCTGTCATCGGTTACTCAGATTTAGAAAGGG GATTTTCGATGTTGCTGTTTTGCCATTTTCCAACAGCACAAATAA CGGGTTATTGTTTATAAATACTACTATTGCCAGCATTGCTGCTA AAGAAGAAGGGGTATCTCTCGAGAAAAGAGAGGCTGAAGCTT ATGTCGAGTTCGCCGGAAGAAGAAGAAGTGTTCATGGTGCG CCGTTAGTCAACCTGAGGCTACAAAGTGTTTTCAATGGCAGAG AAATATGAGAAAGGTTAGAGGTCCACCTGTTTCTTGTATCAAG AGAGATTCTCCAATCCAATGTATTCAAGCTATTGCTGAGAACA GAGCTGATGCTGTTACTTTGGATGGTGGTTTTATCTACGAAGCT GGTTTGGCTCCATATAAACTTAGACCAGTTGCTGCTGAGGTTT ACGGTACTGAAAGACAACCTAGAACTCATTACTATGCTGTTGC TGTGTTAAGAAAGGTGGTTCTTTCCAATTGAACGAATTGCAA GGTTTGAAGTCTTGTCACACTGGTTTGAGAAGAACTGCTGGTT GGAATGTTCCAATTGGTACTTTAAGACCATTCTTAAGTGGACT GGTCCACCTGAGCCAATTGAAGCTGCTGTTGCTAGATTTTTCTC TGCTTCTTGTTCCAGGTGCTGATAAGGGTCAATTCCTAATT TGTGTAGATTGTGTGCTGGTACTGGAGAGAACAATGTGCTTT CTCTTCTCAAGAACCTTACTTTTCTTATTCTGGTGCTTTCAAGT GTTTGAGAGATGGTGCTGGAGATGTTGCTTTTATTAGAGAGTC TACTGTTTTCGAAGATTTGTCTGATGAGGCTGAAAGAGATGAG TATGAATTGTTGTGTCCAGATAACACTAGAAAGCCTGTTGATA AGTTTAAAGATTGTCATTTGGCTAGAGTTCCATCTCACGCTGTT GTTGCTAGATCTGTTAATGGTAAAGAGGATGCTATTTGGAAGT TGTGAGACAAGCTCAAGAAAAGTTCGGTAAAGACAAGTCTCC AAAGTTCCAATTGTTTCGGTTCTCCTTCTGGTCAAAGGATTTGT TGTTTAAAGATTCTGCTATCGGTTTCTCTAGAGTTCCACCTAGA ATTGATTCTGGTTTGTACTTGGGTTCTGGTTACTTCACTGCTAT CCAAAATTTGAGAAAGTCTGAAGAGGAAGTTGCTGCTAGAAG AGCTAGAGTTGTTTGGTGTGCTGTTGGAGAGCAAGAATTGAGA AAGTGTAACCAATGGTCTGGTTTGTCTGAAGGTTCTGTTACTTG TTCTTCTGCTTCTACTACTGAGGATTGTATTGCTTTGGTTTTGAA AGGTGAAGCTGATGCTATGTCTTTGGATGGTGGTTACGTTTATA CTGCTGGTAAATGTGGTTTGGTTCCAGTTTTGGCTGAGAATTAC AAATCTCAACAATCTTCTGATCCAGATCCTAACTGTGTTGATA	46

(continued)

Description	Sequence	SEQ ID NO:
	GACCTGTTGAAGGTTATTTGGCTGTTGCTGTTGTTAGAAGATCT GATACTTCTTTGACTTGGAACCTCTGTTAAAGGTAAAAAGTCTTG TCATACTGCTGTTGATAGAACCGCCGGTTGGAATATTC CAATG GGT TTGTTGTTTAACCAA ACTGGTTCTTGTAAGTTTGATGAGTA CTTCTCTCAATCTTGTGCTCCAGGTTCTGATCCTAGATCTAATT TGTGTGCTTTGTGTATTGGAGATGAGCAAGGTGAAAACAAATG TGTTCCTAATTCTAACGAGAGATACTATGGTTATACTGGTGCTT TTAGATGTTTGGCTGAAAACGCCGGAGATGTTGCTTTCGTTAA GGATGTTACTGTTTTGCAAAACACTGATGGTAACAATAACGAA GCTTGGGCTAAGGATTTGAAATTGGCTGATTTTCGCTTTGTTGTG TTTGGATGGTAAAAGAAAACCAGTTACTGAGGCTAGATCTTGT CATTTGGCTATGGCTCCTAACCACGCTGTTGTTTCTAGAATGGA TAAGGTTGAAAGATTGAAGCAAGTTTTGTTGCATCAACAGGCT AAGTTTGGTAGAAATGGTTCTGATTGTCCTGATAAGTTTTGTTT GTTCCAATCTGAGACTAAAACTTGTTGTTCAATGATAACACT GAATGTTTGGCTAGATTGCACGGTAAAACTACTTACGAAAAAT ATTTGGGTCCTCAATACGTTGCTGGTATTACTAACTTGAAGAA ATGCTCCACCAGTCCATTGCTTGAGGCTTGCGAGTTCCTTAGA AAATAA	

(continued)

Description	Sequence	SEQ ID NO:
SP2-codon optimized hLF gene	ATGCAGTTTGGAAAGGTTCTATTTGCTATTTCTGCCCTGGCTGT CACAGCTCTGGGAGCTCCAGTCAACACTACAACAGAAGATGA AACGGCACAAATTCGGGCTGAAGCTGTCATCGGTTACTCAGAT TTAGAAGGGGATTTCGATGTTGCTGTTTTGCCATTTTCCAACAG CACAAATAACGGGTTATTGTTTATAAATACTACTATTGCCAGC ATTGCTGCTAAAGAAGAAGGGGTATCTCTCGAGAAAAGAGAG GCTGAAGCTTATGTCGAGTTCGCCGGAAGAAGAAGAAGTGTTT AATGGTGCGCCGTTAGTCAACCTGAGGCTACAAAGTGTTTTCA ATGGCAGAGAAATATGAGAAAGGTTAGAGGTCCACCTGTTTCT TGTATCAAGAGAGATTCTCCAATCCAATGTATTCAAGCTATTG CTGAGAACAGAGCTGATGCTGTTACTTTGGATGGTGGTTTTATC TACGAAGCTGGTTTGGCTCCATATAAACTTAGACCAGTTGCTG CTGAGGTTTACGGTACTGAAAGACAACCTAGAACTCATTACTA TGCTGTTGCTGTTGTTAAGAAAGGTGGTTCTTTCCAATTGAACG AATTGCAAGGTTTGAAGTCTTGTCACACTGGTTTGAGAAGAAC TGCTGGTTGGAATGTTCCAATTGGTACTTTAAGACCATTCTTA ACTGGACTGGTCCACCTGAGCCAATTGAAGCTGCTGTTGCTAG ATTTTTCTCTGCTTCTTGTTCCAGGTGCTGATAAGGGTCAAT TTCCTAATTTGTGTAGATTGTGTGCTGGTACTGGAGAGAACAA ATGTGCTTTCTCTTCTCAAGAACCTTACTTTTCTTATTCTGGTGC TTTCAAGTGTTTGAGAGATGGTGCTGGAGATGTTGCTTTTATTA GAGAGTCTACTGTTTTCGAAGATTTGTCTGATGAGGCTGAAAG AGATGAGTATGAATTGTTGTGTCCAGATAACACTAGAAAGCCT GTTGATAAGTTTAAAGATTGTCATTTGGCTAGAGTTCCATCTCA CGCTGTTGTTGCTAGATCTGTTAATGGTAAAGAGGATGCTATTT GGAACCTTGTTGAGACAAGCTCAAGAAAAGTTCGGTAAAGACA AGTCTCCAAAGTTCCAATTGTTTCGGTTCTCCTTCTGGTCAAAAG GATTTGTTGTTTAAAGATTCTGCTATCGGTTTCTCTAGAGTTCC ACCTAGAATTGATTCTGGTTTGTACTTGGGTTCTGGTTACTTCA CTGCTATCCAAAATTTGAGAAAGTCTGAAGAGGAAGTTGCTGC TAGAAGAGCTAGAGTTGTTTGGTGTGCTGTTGGAGAGCAAGAA TTGAGAAAGTGTAACCAATGGTCTGGTTTGTCTGAAGGTTCTG TTAATTGTTCTTCTGCTTCTACTACTGAGGATTGTATTGCTTTGG TTTTGAAAGGTGAAGCTGATGCTATGTCTTTGGATGGTGGTTAC GTTTATACTGCTGGTAAATGTGGTTTGGTTCCAGTTTGGGCTGA GAATTACAAATCTCAACAATCTTCTGATCCAGATCCTAACTGT	47

(continued)

Description	Sequence	SEQ ID NO:
	GTTGATAGACCTGTTGAAGGTTATTTGGCTGTTGCTGTTGTTAG AAGATCTGATACTTCTTTGACTTGGAACCTCTGTTAAAGGTAAA AAGTCTTGTCATACTGCTGTTGATAGAACCGCCGGTTGGAATA TTCCAATGGGTTTGTGTTTAACCAAACCTGGTTCTTGTAAGTTT GATGAGTACTTCTCTCAATCTTGTGCTCCAGGTTCTGATCCTAG ATCTAATTTGTGTGCTTTGTGTATTGGAGATGAGCAAGGTGAA AACAAATGTGTTCCCTAATTCTAACGAGAGATACTATGGTTATA CTGGTGCTTTTAGATGTTTGGCTGAAAACGCCGGAGATGTTGC TTTCGTTAAGGATGTTACTGTTTGCAAAACACTGATGGTAACA ATAACGAAGCTTGGGCTAAGGATTTGAAATTGGCTGATTTTCGC TTTGTTGTGTTTGGATGGTAAAAGAAAACCAAGTTACTGAGGCT AGATCTTGTCATTTGGCTATGGCTCCTAACCACGCTGTTGTTTC TAGAATGGATAAGGTTGAAAGATTGAAGCAAGTTTTGTTGCAT CAACAGGCTAAGTTTGGTAGAAATGGTTCTGATTGTCCTGATA AGTTTTGTTTGTCCAATCTGAGACTAAAAACTTGTTGTTCAAT GATAACACTGAATGTTTGGCTAGATTGCACGGTAAAACCTACTT ACGAAAAATATTTGGGTCCTCAATACGTTGCTGGTATTACTAA CTTGAAGAAATGCTCCACCAGTCCATTGCTTGAGGCTTGCGAG TTCCTTAGAAAATAA	



(continued)

Description	Sequence	SEQ ID NO:
SP3-codon optimized hLF gene	ATGAAATTCATCTCAATTCTGTTTCCTTTTGATAGGCAGTGTATT TGGTGCTCCAGTTGCTCCAGCCGAAGAGGCAGCAAACCACTTG CACAAGCGTGCCGGAAGAAGAAGAAGTGTTC AATGGTGCGCC GTTAGTCAACCTGAGGCTACAAAGTGT TTTCAATGGCAGAGAA ATATGAGAAAGGTTAGAGGTCCACCTGTTTCTTGTATCAAGAG AGATTCTCCAATCCAATGTATTCAAGCTATTGCTGAGAACAGA GCTGATGCTGTTACTTTGGATGGTGGTTTTATCTACGAAGCTGG TTTGGCTCCATATAAACTTAGACCAGTTGCTGCTGAGGTTTACG G TACTGAAAGACAACCTAGAACTCATTACTATGCTGTTGCTGT TGTTAAGAAAGGTGGTTC TTTCCAATTGAACGAATTGCAAGGT TTGAAGTCTTGTCACTGGTTTGAGAAGAACTGCTGGTTGGA ATGTTCCAATTGGTACTTTAAGACCATTTCTTAACTGGACTGGT CCACCTGAGCCAATTGAAGCTGCTGTTGCTAGATTTTTCTCTGC TTCTTGTTGTTCCAGGTGCTGATAAGGGTCAATTTCTTAATTTGT GTAGATTGTGTGCTGGTACTGGAGAGAACAAATGTGCTTTCTC TTCTCAAGAACCTTACTTTTCTTATTCTGGTGCTTTCAAGTGTT GAGAGATGGTGCTGGAGATGTTGCTTTTATTAGAGAGTCTACT GTTTTCGAAGATTTGTCTGATGAGGCTGAAAGAGATGAGTATG AATTGTTGTGTCCAGATAACACTAGAAAGCCTGTTGATAAGTT TAAAGATTGTCATTTGGCTAGAGTTCCATCTCACGCTGTTGTTG CTAGATCTGTTAATGGTAAAGAGGATGCTATTTGGAAC TTGTT GAGACAAGCTCAAGAAAAGTTCGGTAAAGACAAGTCTCCAAA GTTCCAATTGTTTCGGTTCTCCTTCTGGTCAAAAGGATTTGTTGT TTAAAGATTCTGCTATCGGTTTCTCTAGAGTTCCACCTAGAATT GATTCTGGTTTGTACTTGGGTTCTGGTTACTTCACTGCTATCCA AAATTTGAGAAAGTCTGAAGAGGAAGTTGCTGCTAGAAGAGC TAGAGTTGTTTGGTGTGCTGTTGGAGAGCAAGAATTGAGAAAG TGTAACCAATGGTCTGGTTTGTCTGAAGGTTCTGTTACTTGTTC TTCTGCTTCTACTACTGAGGATTGTATTGCTTTGGTTTTGAAAG GTGAAGCTGATGCTATGCTTTTGGATGGTGGTTACGTTTATACT GCTGGTAAATGTGGTTTGGTTCCAGTTTTGGCTGAGAATTACA AATCTCAACAATCTTCTGATCCAGATCCTAACTGTGTTGATAG ACCTGTTGAAGGTTATTTGGCTGTTGCTGTTGTTAGAAGATCTG A TACTTCTTTGACTTGGA ACTCTGT TAAAGGTAAAAAGTCTTGT CATACTGCTGTTGATAGAACCGCCGGTTGGAATATTCCAATGG GTTTGTGTTTAACCAAACTGGTTC TTGTAAGTTTGATGAGTAC	48

(continued)

Description	Sequence	SEQ ID NO:
	TTCTCTCAATCTTGTGCTCCAGGTTCTGATCCTAGATCTAATTT GTGTGCTTTGTGTATTGGAGATGAGCAAGGTGAAAACAAATGT GTTCCTAATTCTAACGAGAGATACTATGGTTATACTGGTGCTTT TAGATGTTTGGCTGAAAACGCCGGAGATGTTGCTTTCGTTAAG GATGTTACTGTTTTGCAAAACACTGATGGTAACAATAACGAAG CTTGGGCTAAGGATTTGAAATTGGCTGATTCGCTTTGTTGTGT TTGGATGGTAAAAGAAAACCAGTACTGAGGCTAGATCTTGTC ATTTGGCTATGGCTCCTAACCACGCTGTTGTTTCTAGAATGGAT AAGGTTGAAAGATTGAAGCAAGTTTTGTTGCATCAACAGGCTA AGTTTGGTAGAAATGGTTCTGATTGTCCTGATAAGTTTTGTTTG TTCCAATCTGAGACTAAAACTTGTTGTTCAATGATAAACTG AATGTTTGGCTAGATTGCACGGTAAACTACTTACGAAAAATA TTTGGGTCCTCAATACGTTGCTGGTATTACTAACTTGAAGAAAT GCTCCACCAGTCCATTGCTTGAGGCTTGCGAGTTCCTTAGAAA ATAA	

(continued)

Description	Sequence	SEQ ID NO:
SP4-codon optimized hLF gene	ATGCAGTTTGGAAAGGTTCTATTTGCTATTTCTGCCCTGGCTGT CACAGCTCTGGGAGCTCCAGTTGCTCCAGCCGAAGAGGCAGCA AACCACTTGCACAAGCGTGCCGGAAGAAGAAGAAGTGTTCAA TGGTGCGCCGTTAGTCAACCTGAGGCTACAAAGTGTTTTCAAT GGCAGAGAAATATGAGAAAGGTTAGAGGTCCACCTGTTTCTTG TATCAAGAGAGATTCTCCAATCCAATGTATTCAAGCTATTGCT GAGAACAGAGCTGATGCTGTTACTTTGGATGGTGGTTTTATCT ACGAAGCTGGTTTGGCTCCATATAAACTTAGACCAGTTGCTGC TGAGGTTTACGGTACTGAAAGACAACCTAGAACTCATTACTAT GCTGTTGCTGTTGTTAAGAAAGGTGGTTCTTTCCAATTGAACGA ATTGCAAGGTTTGAAGTCTTGTCACTGGTTTGAGAAGAAGT GCTGGTTGGAATGTTCCAATTGGTACTTTAAGACCATTTCTTAA CTGGACTGGTCCACCTGAGCCAATTGAAGCTGCTGTTGCTAGA TTTTCTCTGCTTCTTGTGTTCCAGGTGCTGATAAGGGTCAATT TCCTAATTTGTGTAGATTGTGTGCTGGTACTGGAGAGAACAAA TGTGCTTTCTCTTCTCAAGAACCTTACTTTTCTTATTCTGGTGCT TTCAAGTGTTTGAGAGATGGTGCTGGAGATGTTGCTTTTATTAG AGAGTCTACTGTTTTCGAAGATTTGTCTGATGAGGCTGAAAGA GATGAGTATGAATTGTTGTGTCCAGATAACACTAGAAAGCCTG TTGATAAGTTTAAAGATTGTCATTTGGCTAGAGTTCCATCTCAC GCTGTTGTTGCTAGATCTGTTAATGGTAAAGAGGATGCTATTTG GAACTTGTTGAGACAAGCTCAAGAAAAGTTCGGTAAAGACAA GTCTCCAAAGTTCCAATTGTTTCGGTTCTCCTTCTGGTCAAAAGG ATTTGTTGTTTAAAGATTCTGCTATCGGTTTCTCTAGAGTTCCA CCTAGAATTGATTCTGGTTTGTACTTGGGTTCTGGTTACTTCAC TGCTATCCAAAATTTGAGAAAGTCTGAAGAGGAAGTTGCTGCT AGAAGAGCTAGAGTTGTTTGGTGTGCTGTTGGAGAGCAAGAAT TGAGAAAGTGTAACCAATGGTCTGGTTTGTCTGAAGGTTCTGT TACTTGTTCTTCTGCTTCTACTACTGAGGATTGTATTGCTTTGGT TTTGAAAGGTGAAGCTGATGCTATGTCTTTGGATGGTGGTTAC GTTTATACTGCTGGTAAATGTGGTTTGGTTCCAGTTTTGGCTGA GAATTACAAATCTCAACAATCTTCTGATCCAGATCCTAACTGT GTTGATAGACCTGTTGAAGGTTATTTGGCTGTTGCTGTTGTTAG AAGATCTGATACTTCTTTGACTTGGAAGTCTGTTAAAGGTAAG AAGTCTTGTCATACTGCTGTTGATAGAACCGCCGGTTGGAATA TTCCAATGGGTTTGTGTTTAAACCAAACTGGTTCTTGTAAGTTT GATGAGTACTTCTCTCAATCTTGTGCTCCAGGTTCTGATCCTAG ATCTAATTTGTGTGCTTTGTGTATTGGAGATGAGCAAGGTGAA AACAAATGTGTTCTTAATTCTAACGAGAGATACTATGGTTATA CTGGTGCTTTTAGATGTTTGGCTGAAAACGCCGGAGATGTTGC	49

(continued)

Description	Sequence	SEQ ID NO:
	TTTCGTTAAGGATGTTACTGTTTTGCAAAACACTGATGGTAACA ATAACGAAGCTTGGGCTAAGGATTTGAAATTGGCTGATTTTCGC TTTGTGTGTTTGGATGGTAAAAGAAAACCAGTTACTGAGGCT AGATCTTGTCAATTTGGCTATGGCTCCTAACCACGCTGTTGTTTC TAGAATGGATAAGGTTGAAAGATTGAAGCAAGTTTTGTTGCAT CAACAGGCTAAGTTTGGTAGAAATGGTTCTGATTGTCCTGATA AGTTTTGTTTGTTCATCTGAGACTAAAACTTGTGTTCAAT GATAACACTGAATGTTTGGCTAGATTGCACGGTAAAACTACTT ACGAAAAATATTTGGGTCCTCAATACGTTGCTGGTATTACTAA CTTGAAGAAATGCTCCACCAGTCCATTGCTTGAGGCTTGCAG TTCCTTAGAAAATAA	

#### IV. Genetic Engineering

**[0084]** Vectors for transforming microorganisms (e.g., fungal cells, yeast cells) in accordance with the present disclosure can be prepared by known techniques familiar to those skilled in the art in view of the disclosure herein. A vector typically contains one or more genes, in which each gene codes for the expression of a desired product (the gene product) and is operably linked to one or more control sequences that regulate gene expression or target the gene product to a particular location in the recombinant cell.

**[0085]** Exogenous nucleic acid sequences, including, for example, nucleic acid sequences encoding fusion proteins, nucleic acid sequences encoding wild-type or mutant proteins, may be introduced into many different host cells. Nucleic acid sequences configured to facilitate a genetic mutation in a gene may also be introduced into various host cells, as described further herein. Suitable host cells are microbial hosts that can be found broadly within the fungal families. Examples of suitable host strains include but are not limited to fungal or yeast species, such as *Arxula*, *Aspegillus*, *Aurantiochytrium*, *Candida*, *Claviceps*, *Cryptococcus*, *Cunninghamella*, *Hansenula*, *Kluyveromyces*, *Komagataella*, *Leucosporidiella*, *Lipomyces*, *Mortierella*, *Ogataea*, *Pichia*, *Prototheca*, *Rhizopus*, *Rhodospordium*, *Rhodotorula*, *Saccharomyces*, *Schizosaccharomyces*, *Tremella*, *Trichosporon*, and *Yarrowia*. In some embodiments, a host cell of the present disclosure is a *Komagataella* cell. In some embodiments, a host cell of the present disclosure is *Komagataella phaffii*. In some embodiments, a host cell of the present disclosure is *Komagataella pastoris*. In some embodiments, a host cell of the present disclosure is *Komagataella pseudopastoris*.

**[0086]** Microbial expression systems and expression vectors are well known to those skilled in the art. Any such expression vector could be used to introduce the instant genes and nucleic acid sequences into an organism. The nucleic acid sequences may be introduced into appropriate microorganisms via transformation techniques. For example, a nucleic acid sequence can be cloned in a suitable plasmid, and a parent cell can be transformed with the resulting plasmid. The plasmid is not particularly limited so long as it renders a desired nucleic acid sequence inheritable to the microorganism's progeny.

**[0087]** Vectors or cassettes useful for the transformation of suitable host cells are recognized in the art. Typically the vector or cassette contains a gene, sequences directing transcription and translation of a relevant gene including the promoter, a selectable marker, and sequences allowing autonomous replication or chromosomal integration. Suitable vectors comprise a region 5' of the gene harboring the promoter and other transcriptional initiation controls and a region 3' of the DNA fragment which controls transcriptional termination.

**[0088]** Promoters, cDNAs, and 3'UTRs, as well as other elements of the vectors, can be generated through cloning techniques using fragments isolated from native sources (Green & Sambrook, *Molecular Cloning: A Laboratory Manual*, (4th ed., 2012); U.S. Pat. No. 4,683,202; incorporated by reference). Alternatively, elements can be generated synthetically using known methods (Gene 164:49-53 (1995)).

#### A. Vectors and Vector Components

**[0089]** Vectors for transforming microorganisms (e.g., yeast cells) in accordance with the present disclosure can be prepared by known techniques familiar to those skilled in the art in view of the disclosure herein. A vector typically contains one or more genes, in which each gene codes for the expression of a desired product (the gene product) and

is operably linked to one or more control sequences (e.g., promoter sequences, signal peptide sequences) that regulate gene expression or target the gene product to a particular location in the recombinant cell.

## 1. Control Sequences

**[0090]** Control sequences are nucleic acid sequences that regulate the expression of a coding sequence or direct a gene product to a particular location in or outside a cell. Control sequences that regulate expression include, for example, promoters that regulate transcription of a coding sequence and terminators that terminate transcription of a coding sequence. Another control sequence is a 3' untranslated sequence located at the end of a coding sequence that encodes a polyadenylation signal. Control sequences that direct gene products to particular locations include those that encode signal peptides, which direct the protein to which they are attached to a particular location inside or outside the cell.

**[0091]** Thus, an example vector design for expression of a gene in a microbe contains a coding sequence for a desired gene product (for example, a selectable marker, an enzyme, a fusion protein, etc.) in operable linkage with a promoter active in yeast. Alternatively, if the vector does not contain a promoter in operable linkage with the coding sequence of interest, the coding sequence can be transformed into the cells such that it becomes operably linked to an endogenous promoter at the point of vector integration. Example promoters contemplated herein include, but are not limited to, the AOX1, GAP, TEF1, TPI1, DAS1, DAS2, CAT1, and FMD promoters.

**[0092]** The promoter used to express a gene can be the promoter naturally linked to that gene or a different promoter.

**[0093]** A promoter can generally be characterized as constitutive or inducible. Constitutive promoters are generally active or function to drive expression at all times (or at certain times in the cell life cycle) at the same level. Inducible promoters, conversely, are active (or rendered inactive) or are significantly up- or down-regulated only in response to a stimulus. Both types of promoters find application in the disclosed methods. Useful inducible promoters include those that mediate transcription of an operably linked gene in response to a stimulus, such as an exogenously provided small molecule, temperature (heat or cold), lack of nitrogen in culture media, etc. Suitable promoters can activate transcription of an essentially silent gene or upregulate transcription of an operably linked gene that is transcribed at a low level.

**[0094]** Inclusion of termination region control sequence is optional. The termination region may be native to the transcriptional initiation region (the promoter), may be native to the DNA sequence of interest, or may be obtainable from another source (See, e.g., Chen & Orozco, *Nucleic Acids Research* 16:8411 (1988)).

**[0095]** In some cases, the full nucleotide sequence of a promoter is not necessary to drive transcription, and sequences shorter than the promoter's full nucleotide sequence can drive transcription of an operably-linked gene. The minimal portion of a promoter, termed the core promoter, includes a transcription start site, a binding site for a RNA polymerase, and a binding site for a transcription factor.

**[0096]** A promoter may be linked to a target by introducing the promoter and the target into a nucleic acid molecule, for example, a vector. A vector may be introduced into a cell, thereby expressing the promoter and the target. In one embodiment, a promoter is linked to a target by introducing a promoter into DNA of a cell, for example, via homologous recombination, thereby integrating the promoter into the genome of the cell.

## B. Genes and Codon Optimization

**[0097]** Typically, a gene includes a promoter, a coding sequence, and termination control sequences. When assembled by recombinant DNA technology, a gene may be termed an expression cassette and may be flanked by restriction sites for convenient insertion into a vector that is used to introduce the recombinant gene into a host cell. The expression cassette can be flanked by DNA sequences from the genome or other nucleic acid target to facilitate stable integration of the expression cassette into the genome by homologous recombination. Alternatively, the vector and its expression cassette may remain unintegrated (e.g., an episome), in which case, the vector typically includes an origin of replication, which is capable of providing for replication of the vector DNA.

**[0098]** A common gene present on a vector is a gene that codes for a protein, the expression of which allows the recombinant cell containing the protein to be differentiated from cells that do not express the protein. Such a gene, and its corresponding gene product, is called a selectable marker or selection marker. Any of a wide variety of selectable markers can be employed in a transgene construct useful for transforming the organisms covered in the disclosed embodiments.

**[0099]** For optimal expression of a recombinant protein, it may be beneficial to employ coding sequences that produce mRNA with codons optimally used by the host cell to be transformed. Thus, proper expression of transgenes can require that the codon usage of the transgene matches the specific codon bias of the organism in which the transgene is being expressed. The precise mechanisms underlying this effect are many, but include the proper balancing of available aminoacylated tRNA pools with proteins being synthesized in the cell, coupled with more efficient translation of the transgenic messenger RNA (mRNA) when this need is met. When codon usage in the transgene is not optimized, available tRNA pools may not be sufficient to allow for efficient translation of the transgenic mRNA resulting in ribosomal

stalling and termination and possible instability of the transgenic mRNA.

**[0100]** A coding sequence of the present disclosure can be codon optimized for a particular host cell by replacing one or more rare codons with one or more codons more frequently found in the host cell. A rare codon in a host cell describes a codon that is found in less than 5%, less than 10%, or less than 20% of coding sequences in the host cell. Rare codons

can be identified using methods known to those of skill in the art.

**[0101]** Aspects of the disclosure comprise transformation of a microorganism with a nucleic acid sequence comprising a gene that encodes a protein. The gene may be native to the cell or from a different species. The gene may be derived from a different species yet modified (e.g., codon optimized) for optimal expression in the microorganism. In certain embodiments, the gene is inheritable to the progeny of a transformed cell. In some embodiments, the gene is inheritable because it resides on a plasmid. In certain embodiments, the gene is inheritable because it is integrated into the genome of the transformed cell.

**[0102]** Further aspects of the disclosure may comprise transformation of a microorganism with a nucleic acid sequence configured to generate a mutation in a gene of the microorganism. For example, aspects of the disclosure may comprise transformation of the microorganism with a nucleic acid sequence comprising sequences upstream and downstream of a gene (e.g., an OCH1 gene), thereby facilitating reduced expression or deletion of the gene via homologous recombination. Various methods for generating mutations (including deletions or knockout mutations, as well as mutations which reduce expression of a gene) in genes of a microorganism are recognized in the art and envisioned herein. A microorganism having a deletion or knockout mutation of a gene does not product a functional copy of the protein. For example, a recombinant yeast cell of the disclosure may comprise a deletion of an endogenous OCH1 gene, such that the recombinant yeast cell does not express an endogenous, functional OCH1 protein. A microorganism having a reduced expression of a gene or protein produces a functional copy of the protein, but at a reduced amount compared with a wild-type (i.e., a non-recombinant or non-genetically modified) microorganism of the same species. Methods for reducing expression of a protein are recognized in the art and include, for example, replacement of an endogenous promoter and/or modification of one or more regulatory elements.

### C. Transformation

**[0103]** Cells can be transformed by any suitable technique including, e.g., biolistics, electroporation, glass bead transformation, and silicon carbide whisker transformation. Any convenient technique for introducing a transgene into a microorganism can be employed in the embodiments disclosed herein.

**[0104]** Vectors for transformation of microorganisms can be prepared by known techniques familiar to those skilled in the art. In one embodiment, an exemplary vector design for expression of a gene in a microorganism contains a gene encoding an enzyme in operable linkage with a promoter active in the microorganism. Alternatively, if the vector does not contain a promoter in operable linkage with the gene of interest, the gene can be transformed into the cells such that it becomes operably linked to a native promoter at the point of vector integration. The vector can also contain a second gene that encodes a protein. Optionally, one or both gene(s) is/are followed by a 3' untranslated sequence containing a polyadenylation signal. Expression cassettes encoding the two genes can be physically linked in the vector or on separate vectors. Co-transformation of microbes can also be used, in which distinct vector molecules are simultaneously used to transform cells (Protist 155:381-93 (2004)). The transformed cells can be optionally selected based upon the ability to grow in the presence of the antibiotic or other selectable marker under conditions in which cells lacking the resistance cassette would not grow.

### D. Genetically Engineered Cells

**[0105]** Aspects of the disclosure comprise genetically engineered cells (also "engineered cells" or "recombinant cells") and methods for making and using such cells. In some embodiments, disclosed are recombinant cells comprising one or more exogenous nucleic acid sequences. Also disclosed are methods for generating such recombinant cells comprising introducing the one or more exogenous nucleic acid sequences into a host cell. Further described are methods for collecting one or more products (e.g., a mammalian protein) from such recombinant cells comprising culturing the cells and collecting the product.

**[0106]** In some embodiments, the recombinant cell is a prokaryotic cell, such as a bacterial cell. In some embodiments, the recombinant cell is a eukaryotic cell, such as a mammalian cell, a yeast cell, a filamentous fungi cell, a protist cell, an algae cell, an avian cell, a plant cell, or an insect cell. In some embodiments, the cell is a yeast cell. Those with skill in the art will recognize that many forms of filamentous fungi produce yeast-like growth, and the definition of yeast herein encompasses such cells. A recombinant cell of the disclosure may be selected from the group consisting of algae, bacteria, molds, fungi, plants, and yeasts. In some embodiments, a recombinant cell of the disclosure is a bacterial cell (e.g. *E.coli*), a fungal cell, or a yeast cell.

**[0107]** In some embodiments, a recombinant cell of the disclosure is a recombinant fungal cell. A recombinant fungal

cell may be any suitable fungal cell recognized in the art. In some aspects, the fungal cell is an *Arxula*, *Aspegillus*, *Aurantiochytrium*, *Candida*, *Claviceps*, *Cryptococcus*, *Cunninghamella*, *Geotrichum*, *Hansenula*, *Kluyveromyces*, *Kodamaea*, *Komagataella*, *Leucosporidiella*, *Lipomyces*, *Mortierella*, *Ogataea*, *Pichia*, *Prototheca*, *Rhizopus*, *Rhodospiridium*, *Rhodotorula*, *Saccharomyces*, *Schizosaccharomyces*, *Tremella*, *Trichosporon*, *Wickerhamomyces*, or *Yarrowia* cell. In some embodiments, the fungal cell is *Arxula adeninivorans*, *Aspergillus niger*, *Aspergillus oryzae*, *Aspergillus terreus*, *Aurantiochytrium limacinum*, *Candida utilis*, *Claviceps purpurea*, *Cryptococcus albidus*, *Cryptococcus curvatus*, *Cryptococcus ramirezgomezianus*, *Cryptococcus terreus*, *Cryptococcus wieringae*, *Cunninghamella echinulata*, *Cunninghamella japonica*, *Geotrichum fermentans*, *Hansenula polymorpha*, *Kluyveromyces lactis*, *Komagataella phaffii*, *Komagataella pastoris*, *Komagataella pseudopastoris*, *Kluyveromyces marxianus*, *Kodamaea ohmeri*, *Leucosporidiella creatinivora*, *Lipomyces lipofer*, *Lipomyces starkeyi*, *Lipomyces tetrasporus*, *Mortierella isabellina*, *Mortierella alpina*, *Ogataea polymorpha*, *Pichia ciferrii*, *Pichia guilliermondii*, *Pichia pastoris*, *Pichia stipites*, *Prototheca zopfii*, *Rhizopus arrhizus*, *Rhodospiridium babjevae*, *Rhodospiridium toruloides*, *Rhodospiridium paludigenum*, *Rhodotorula glutinis*, *Rhodotorula mucilaginosa*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Tremella encephala*, *Trichosporon cutaneum*, *Trichosporon fermentans*, *Wickerhamomyces ciferrii*, or *Yarrowia lipolytica*.

**[0108]** In some aspects, the fungal cell is a yeast cell. In certain embodiments, the yeast cell is a *Komagataella* cell. In some embodiments, the yeast cell is *Kluyveromyces phaffii*, *Komagataella pastoris*, or *Komagataella pseudopastoris*. In particular embodiments, the yeast cell is *Kluyveromyces phaffii*.

**[0109]** In some embodiments, an engineered cell of the present disclosure is a yeast cell comprising one or more modifications for improving generation of N-glycans including human-like N-glycans. Examples of such cells and modifications are described in, for example, US patent 9,617,550, incorporated herein by reference in its entirety.

## E. Gene Editing Systems

**[0110]** Certain embodiments of the disclosure are directed to the use of gene editing techniques to generate a knockout or other mutation in a gene in a population of cells. Various methods and systems for gene editing are known in the art and include, for example, zinc finger nuclease (ZFN)-based gene editing, transcription activator-like effector nuclease (TALEN)-based gene editing, and CRISPR/Cas-based gene editing. Various methods and systems for gene editing are recognized in the art and contemplated herein. In some embodiments, methods of the present disclosure comprise CRISPR/Cas-based gene editing, which comprises the use of components of a CRISPR system, for example a guide RNA (gRNA) and a Cas nuclease. In some embodiments, a method of the present disclosure does not comprise CRISPR/Cas-based gene editing (e.g., comprises ZFN-based, TALEN-based, or any other gene editing method or system).

**[0111]** In general, "CRISPR system" refers collectively to transcripts and other elements involved in the expression of or directing the activity of CRISPR-associated ("Cas") genes, including sequences encoding a Cas gene, a tracr (trans-activating CRISPR) sequence (e.g. tracrRNA or an active partial tracrRNA), a tracr-mate sequence (encompassing a "direct repeat" and a tracrRNA-processed partial direct repeat in the context of an endogenous CRISPR system), a guide sequence (also referred to as a "spacer" in the context of an endogenous CRISPR system), and/or other sequences and transcripts from a CRISPR locus.

**[0112]** The CRISPR/Cas nuclease or CRISPR/Cas nuclease system can include a noncoding RNA molecule (guide) RNA, which sequence-specifically binds to DNA, and a Cas protein (e.g., Cas9), with nuclease functionality (e.g., two nuclease domains). One or more elements of a CRISPR system can derive from a type I, type II, or type III CRISPR system, e.g., derived from a particular organism comprising an endogenous CRISPR system, such as *Streptococcus pyogenes*.

**[0113]** In some aspects, a Cas nuclease and gRNA (including a fusion of crRNA specific for the target sequence and fixed tracrRNA) are introduced into the cell. A Cas nuclease and a gRNA can be introduced into the cell indirectly via introduction of one or more nucleic acids (e.g., vectors) encoding for the Cas nuclease and/or the gRNA. A Cas nuclease and a gRNA can be introduced into the cell directly by introduction of a Cas nuclease protein and a gRNA molecule. In general, target sites at the 5' end of the gRNA target the Cas nuclease to the target site, e.g., the gene, using complementary base pairing. The target site may be selected based on its location immediately 5' of a protospacer adjacent motif (PAM) sequence, such as typically NGG, or NAG. In this respect, the gRNA may be targeted to the desired sequence by modifying the first 20, 19, 18, 17, 16, 15, 14, 13, 12, 11, or 10 nucleotides of the guide RNA to correspond to the target DNA sequence. In general, a CRISPR system is characterized by elements that promote the formation of a CRISPR complex at the site of a target sequence. Typically, "target sequence" generally refers to a sequence to which a guide sequence is designed to have complementarity, where hybridization between the target sequence and a guide sequence promotes the formation of a CRISPR complex. Full complementarity is not necessarily required, provided there is sufficient complementarity to cause hybridization and promote formation of a CRISPR complex.

**[0114]** The CRISPR system can induce double stranded breaks (DSBs) at the target site, followed by disruptions as discussed herein. In other embodiments, Cas9 variants, deemed "nickases," are used to nick a single strand at the target

site. Paired nickases can be used, e.g., to improve specificity, each directed by a pair of different gRNAs targeting sequences such that upon introduction of the nicks simultaneously, a 5' overhang is introduced. In other embodiments, catalytically inactive Cas9 is fused to a heterologous effector domain such as a transcriptional repressor or activator, to affect gene expression.

**[0115]** The target sequence may comprise any polynucleotide, such as DNA or RNA polynucleotides. The target sequence may be located in the nucleus or cytoplasm of the cell, such as within an organelle of the cell. Generally, a sequence or template that may be used for recombination into the targeted locus comprising the target sequences is referred to as an "editing template" or "editing polynucleotide" or "editing sequence". In some aspects, an exogenous template polynucleotide may be referred to as an editing template. In some aspects, the recombination is homologous recombination.

**[0116]** Typically, in the context of an endogenous CRISPR system, formation of the CRISPR complex (comprising the guide sequence hybridized to the target sequence and complexed with one or more Cas proteins) results in cleavage of one or both strands in or near (e.g. within 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 50, or more base pairs from) the target sequence. The tracr sequence, which may comprise or consist of all or a portion of a wild-type tracr sequence (e.g. about or more than about 20, 26, 32, 45, 48, 54, 63, 67, 85, or more nucleotides of a wild-type tracr sequence), may also form part of the CRISPR complex, such as by hybridization along at least a portion of the tracr sequence to all or a portion of a tracr mate sequence that is operably linked to the guide sequence. The tracr sequence has sufficient complementarity to a tracr mate sequence to hybridize and participate in formation of the CRISPR complex, such as at least 50%, 60%, 70%, 80%, 90%, 95% or 99% sequence complementarity along the length of the tracr mate sequence when optimally aligned.

**[0117]** One or more vectors driving expression of one or more elements of a CRISPR system can be introduced into a cell such that expression of the elements of the CRISPR system direct formation of a CRISPR complex at one or more target sites. Components can also be delivered to cells as proteins and/or RNA. For example, a Cas enzyme, a guide sequence linked to a tracr-mate sequence, and a tracr sequence could each be operably linked to separate regulatory elements on separate vectors. Alternatively, two or more of the elements expressed from the same or different regulatory elements, may be combined in a single vector, with one or more additional vectors providing any components of the CRISPR system not included in the first vector. The vector may comprise one or more insertion sites, such as a restriction endonuclease recognition sequence (also referred to as a "cloning site"). In some embodiments, one or more insertion sites are located upstream and/or downstream of one or more sequence elements of one or more vectors. When multiple different guide sequences are used, a single expression construct may be used to target CRISPR activity to multiple different, corresponding target sequences within a cell.

**[0118]** A vector may comprise a regulatory element operably linked to an enzyme-coding sequence encoding a Cas protein (also "Cas nuclease"). Non-limiting examples of Cas proteins include Cas1, Cas1B, Cas2, Cas3, Cas4, Cas5, Cas6, Cas7, Cas8, Cas9 (also known as Csn1 and Csx12), Cas10, Cas12a (Cpf1), Csy1, Csy2, Csy3, Cse1, Cse2, Csc1, Csc2, Csa5, Csn2, Csm2, Csm3, Csm4, Csm5, Csm6, Cmr1, Cmr3, Cmr4, Cmr5, Cmr6, Csb1, Csb2, Csb3, Csx17, Csx14, Csx10, Csx16, CsaX, Csx3, Csx1, Csx15, Csf1, Csf2, Csf3, Csf4, homologs thereof, or modified versions thereof. These enzymes are known; for example, the amino acid sequence of *S. pyogenes* Cas9 protein may be found in the SwissProt database under accession number Q99ZW2.

**[0119]** The Cas nuclease can be Cas9 (e.g., from *S. pyogenes* or *S. pneumoniae*). The Cas nuclease can be Cas12a. The Cas nuclease can direct cleavage of one or both strands at the location of a target sequence, such as within the target sequence and/or within the complement of the target sequence. The vector can encode a Cas nuclease that is mutated with respect to a corresponding wild-type enzyme such that the mutated Cas nuclease lacks the ability to cleave one or both strands of a target polynucleotide containing a target sequence. In some embodiments, a Cas9 nickase may be used in combination with guide sequence(s), e.g., two guide sequences, which target respectively sense and antisense strands of the DNA target. This combination allows both strands to be nicked and used to induce NHEJ or HDR.

**[0120]** In some embodiments, an enzyme coding sequence encoding the CRISPR enzyme is codon optimized for expression in particular cells, such as yeast cells.

**[0121]** In general, a guide sequence is any polynucleotide sequence having sufficient complementarity with a target polynucleotide sequence to hybridize with the target sequence and direct sequence-specific binding of the CRISPR complex to the target sequence. In some embodiments, the degree of complementarity between a guide sequence and its corresponding target sequence, when optimally aligned using a suitable alignment algorithm, is or is more than 50%, 60%, 75%, 80%, 85%, 90%, 95%, 97.5%, 99%, or more.

**[0122]** Optimal alignment may be determined with the use of any suitable algorithm for aligning sequences, non-limiting example of which include the Smith-Waterman algorithm, the Needleman-Wunsch algorithm, algorithms based on the Burrows-Wheeler Transform (e.g. the Burrows Wheeler Aligner), Clustal W, Clustal X, BLAST, Novoalign (Novocraft Technologies, ELAND (Illumina, San Diego, Calif.), SOAP (available at [soap.genomics.org.cn](http://soap.genomics.org.cn)), and Maq (available at [maq.sourceforge.net](http://maq.sourceforge.net)).

**[0123]** The Cas nuclease may be part of a fusion protein comprising one or more heterologous protein domains. A



Cas nuclease fusion protein may comprise any additional protein sequence, and optionally a linker sequence between any two domains. Examples of protein domains that may be fused to a Cas nuclease, without limitation, epitope tags, reporter gene sequences, and protein domains having one or more of the following activities: methylase activity, demethylase activity, transcription activation activity, transcription repression activity, transcription release factor activity, histone modification activity, RNA cleavage activity and nucleic acid binding activity. Non-limiting examples of epitope tags include histidine (His) tags, V5 tags, FLAG tags, influenza hemagglutinin (HA) tags, Myc tags, VSV-G tags, and thioredoxin (Trx) tags. Examples of reporter genes include, but are not limited to, glutathione-S-transferase (GST), horseradish peroxidase (HRP), chloramphenicol acetyltransferase (CAT) beta galactosidase, beta-glucuronidase, luciferase, green fluorescent protein (GFP), HcRed, DsRed, cyan fluorescent protein (CFP), yellow fluorescent protein (YFP), and autofluorescent proteins including blue fluorescent protein (BFP). A Cas nuclease may be fused to a gene sequence encoding a protein or a fragment of a protein that bind DNA molecules or bind other cellular molecules, including but not limited to maltose binding protein (MBP), S-tag, Lex A DNA binding domain (DBD) fusions, GAL4A DNA binding domain fusions, and herpes simplex virus (HSV) BP16 protein fusions. Additional domains that may form part of a fusion protein comprising a Cas nuclease are described in US 20110059502, incorporated herein by reference.

## Examples

**[0124]** The following examples are included to demonstrate certain embodiments disclosed herein. It should be appreciated by those of skill in the art that the techniques disclosed in the examples which follow represent techniques discovered by the inventor to function well in the practice of the disclosed embodiments, and thus can be considered to constitute certain modes for its practice. However, those of skill in the art should, in light of the present disclosure, appreciate that many changes can be made in the specific embodiments which are disclosed and still obtain a like or similar result without departing from the spirit and scope of the embodiments disclosed herein.

### Example 1 - Novel signal peptides increase extracellular protein levels

**[0125]** To determine the effect of the novel signal peptides on extracellular protein levels, DNA encoding SEQ ID NO:1 ("SP1"), SEQ ID NO:2 ("SP2"), and SEQ ID NO:4 ("SP4") was cloned in-frame in the 5' end of the DNA coding for a protein of interest (POI), i.e., Pichia pastoris codon-optimized human lactoferrin, resulting in the substitution of the pre-pro-MF $\alpha$  from Saccharomyces cerevisiae. This is the most widely used signal peptide in yeast and served as the control. Single copies of the resulting sequences and the control were integrated into the AOX1 locus via double-crossover. Multiple colonies of each transformation plate were cultivated in 96-deep well plates.

**[0126]** To establish the presence of the protein of interest, a Western Blot was run on the supernatant. As shown in FIG. 1, when a single copy of human lactoferrin is integrated, and secretion is driven by the widely used pre-pro-MF $\alpha$  from Saccharomyces cerevisiae protein is not detected in the supernatant. Contrarily, extracellular protein is detected when secretion is driven by SEQ ID NO:1 ("SP1"), SEQ ID NO:2 ("SP2"), and SEQ ID NO:3 ("SP3").

**[0127]** With the aim of assessing the magnitude of the improvements in secretion, quantification of extracellular protein was performed via ELISA. As seen in FIG. 2, the novel engineered signals enhanced extracellular protein levels by 2.38-fold, 2.41-fold, and 2.20-fold with respect to the control (pre-pro-MF $\alpha$ ) for SEQ ID NO:1 ("SP1"), SEQ ID NO:2 ("SP2"), and SEQ ID NO:3 ("SP3") respectively.

## Materials and Methods

**[0128]** Vectors and strain construction. Oligonucleotides and gBlocks were ordered from Integrated DNA Technologies (San Diego, CA, USA) and are described in Table 5. NEBuilder<sup>®</sup> HiFi DNA Assembly Master Mix, OneTaq<sup>®</sup> Quickload<sup>®</sup> DNA polymerase, and Escherichia coli DH5 $\alpha$  cells were from New England Biolabs. All polymerase chain reaction (PCR)-amplified sequences were confirmed via sequencing at or Genewiz.

Table 5  
gBLOCK and Primers

SEQ ID NO:	Name	Sequence
------------	------	----------

(continued)

Table 5  
gBLOCK and Primers

5	15	gBLOCK 1	ATGCAGTTTGGAAAGGTTCTATTTGCTATTTCTGCCCTGG CTGTCACAGCTCTGGGAGCTCCAGTTGCTCCAGCCGAAG AGGCAGCAAACCACTTGCACAAGCGTATGAGGCAGGTT 10 TGGTTCTCTTGGATTGTGGGATTGTTTCCTATGTTTTTCA ACGTGTCTTCTGCTAAACGATGAAATTCATCTCAATTCT GTTTCCTTTTGATAGGCAGTGTATTTGGTATGAAATTCATC TCAATTCTGTTTCCTTTTGATAGGCAGTGTATTTGGTGCTC 15 CAGTTGCTCCAGCCGAAGAGGCAGCAAACCACTTGCAC AAGCGT
	16	PMR1	CGAAGGATCCAAACGATGAAATTCATCTCAATTC
	17	PMR2	GTAGTGTTGACTGGAGCACCAATACAC
20	18	PMR3	GGTGCTCCAGTCAACACTACAACAGAAG
	19	PMR4	TTCATCGTTTGGATCCTTCGAATAATTAGTTG
	20	PMR5	CGAAGGATCCAAACGATGCAGTTTGGAAAGG
	21	PMR6	TGACTGGAGCTCCCAGAGCTGTGACAGC
25	22	PMR7	AGCTCTGGGAGCTCCAGTCAACACTACAAC
	23	PMR8	TGCATCGTTTGGATCCTTCGAATAATTAGTTGTTTTTG
	24	PMR9	GATCCAAACGATGAAATTCATCTCAATTCTGTTTCCTTTG
30	25	PMR10	TTCTTCCGGCACGCTTGTGCAAGTGGTTTG
	26	PMR11	GCACAAGCGTGCCGGAAGAAGAAGAAGTG
	27	PMR12	TGAATTTTCATCGTTTGGATCCTTCGAATAATTAG
	28	PMR13	GATCCAAACGATGCAGTTTGGAAAGGTTCTATTTG
35	29	PMR14	TTCTTCCGGCACGCTTGTGCAAGTGGTTTG
	30	PMR15	GCACAAGCGTGCCGGAAGAAGAAGAAGTG
	31	PMR16	CAAACTGCATCGTTTGGATCCTTCGAATAATTAG
40	32	PMR17	GATCTAACATCCAAAGACGAAA
	33	PMR18	TTGAGATAAATTTACGTTTAA

**[0129]** Transformation of linear dsDNA for integration was performed using the method described by Madden, Tolstorukov, & Cregg (2014) Fungi, Volume 1, Fungal Biology. Total yeast genomic DNA extraction was performed using the kit Easy DNA from Invitrogen (ThermoFisher, Applied Biosystems™, PrepSEQ™ 1-2-3 Nucleic Acid Extraction Kit, Catalog number: 4452222). The resulting plasmids are summarized in Table 6.

Table 6  
Plasmids

Name	Description
P1	pPIC9 (Invitrogen) with a codon-optimized version of human lactoferrin lacking its native secretion signal. Secretion is driven by the <i>S. cerevisiae</i> pre-pro-MF $\alpha$ secretion signal.
P2	P1 where the <i>S. cerevisiae</i> pre-pro-MF $\alpha$ secretion signal was substituted SEQ ID NO:1 (ostpro)
P3	P1 where the <i>S. cerevisiae</i> pre-pro-MF $\alpha$ secretion signal was substituted by SEQ ID NO:2

(continued)

Table 6 Plasmids	
Name	Description
P4	P1 where the <i>S. cerevisiae</i> pre-pro-MF $\alpha$ secretion signal was substituted by SEQ ID NO:3
P5	P1 where the <i>S. cerevisiae</i> pre-pro-MF $\alpha$ secretion signal was substituted by SEQ ID NO:4

**[0130]** The leader peptide sequences from the *Pichia pastoris* endogenous proteins Ost1 and Pst1 were determined using SignalP-5.0 bioinformatic software, publicly available from the Center Biological Sequence Analysis (CBS). The pro region of Epx1 was described by Heiss et al. (2015) Microbiology, 161(7).

**[0131]** Plasmid P1 containing the gene encoding human lactoferrin without its native secretion peptide fused in-frame with the pre-pro-leader peptide of the mating factor- $\alpha$  from *Saccharomyces cerevisiae* was synthesized by Genscript. The human lactoferrin gene was codon-optimized for expression in *Pichia pastoris*.

**[0132]** To create plasmid P2 containing signal sequence SP1 (SEQ ID NO:1), primers PMR1 (SEQ ID NO:16) and PMR2 (SEQ ID NO:17) were used to amplify the Ost1 leader sequence using gBLOCK1 as a template. The backbone containing human lactoferrin, yeast *HIS4* auxotrophic marker, and *Escherichia coli* antibiotic resistance and origin of replication was obtained via polymerase chain reaction (PCR) of P1 plasmid using primers PMR3 (SEQ ID NO:18) and PMR4 (SEQ ID NO: 19). The two resulting fragments were assembled using NEBuilder<sup>®</sup> HiFi DNA Assembly Master Mix following manufacturer instructions.

**[0133]** To generate plasmid P3 containing signal sequence SP2 (SEQ ID NO:2), primers PMR5 (SEQ ID NO:20) and PMR6 (SEQ ID NO:21) were used for amplification using gBLOCK1 (SEQ ID NO: 15) as a template. The backbone containing human lactoferrin, yeast *HIS4* auxotrophic marker, and *Escherichia coli* antibiotic resistance and origin of replication was obtained via PCR of P1 plasmid with primers PMR7 (SEQ ID NO:22) and PMR8 (SEQ ID NO:23). The two resulting fragments were assembled using NEBuilder<sup>®</sup> HiFi DNA Assembly Master Mix following manufacturer instructions.

**[0134]** To generate plasmid P4 containing signal sequence SP3 (SEQ ID NO:3), primers PMR9 (SEQ ID NO:24) and PMR10 (SEQ ID NO:25) were used for amplification using the gBLOCK1 as a template. The backbone containing human lactoferrin, yeast *HIS4* auxotrophic marker, and *Escherichia coli* antibiotic resistance and origin of replication was obtained via PCR of P1 plasmid with primers PMR11 (SEQ ID NO:26) and PMR12 (SEQ ID NO:27). The two resulting fragments were assembled using NEBuilder<sup>®</sup> HiFi DNA Assembly Master Mix following manufacturer instructions.

**[0135]** To generate plasmid P5 containing signal sequence SP4 (SEQ ID NO:4), primers PMR13 (SEQ ID NO:28) and PMR14 (SEQ ID NO:29) were used for amplification using the gBLOCK1 (SEQ ID NO: 15) as a template. The backbone containing human lactoferrin, yeast *HIS4* auxotrophic marker, and *Escherichia coli* antibiotic resistance and origin of replication was obtained via PCR of P1 plasmid with primers PMR15 (SEQ ID NO:30) and PMR16 (SEQ ID NO:31). The two resulting fragments were assembled using NEBuilder<sup>®</sup> HiFi DNA Assembly Master Mix following manufacturer instructions.

**[0136]** Assembly mixtures were transformed into *Escherichia coli* DH5 $\alpha$  cells as directed by the manufacturer and plated into Luria Broth (LB)-agar plates containing 100  $\mu$ g/mL of ampicillin. Positive clones were selected via colony polymerase chain reaction (PCR) and inoculated overnight in 5 mL of liquid Luria Broth media supplemented with 100  $\mu$ g/mL of ampicillin. Plasmids from *Escherichia coli* cells were isolated using GeneJET plasmid miniprep kit (ThermoFisher<sup>®</sup>, Catalog number K0502). Proper assembly was confirmed via Sanger DNA sequencing.

**[0137]** Linear dsDNA fragment for integration into yeast was obtained using Q5 High-Fidelity DNA polymerase using primers PMR17 (SEQ ID NO:32) and PMR18 (SEQ ID NO:33) and plasmids P1, P2, P3, P4, or P5 as a template. Electrocompetent *Pichia pastoris* cells were transformed as described by Madden, Tolstorukov, & Cregg (2014) Fungi, Volume 1, Fungal Biology. Cells were spread on MD plates (1.34% yeast nitrogen base,  $4 \times 10^{-5}$  % biotin, 2% dextrose, 20% agar), which allows for selection of *his4*<sup>+</sup> cells, and incubated at 30°C for seventy-two hours. Individual yeast colonies (~10-20) are then re-streaked in MD plates and allowed to grow for twenty-four hours at 30°C. Cells transformed with P1 were used as controls for assessing higher efficiency of SP1 (SEQ ID NO:1), SP2 (SEQ ID NO:2), SP3 (SEQ ID NO:3), and SP4 (SEQ ID NO:5) in the secretion of a protein of interest (POI).

**[0138]** Individual colonies from re-streaked plates are inoculated in 96-deep well plates using 600  $\mu$ l of 2% YPD (2% dextrose, 2% peptone, 1% yeast extract). Cells were grown for forty-eight hours at 1,000 rpm and 30°C. Fifty microliters of the resulting cell suspension were transferred to 550  $\mu$ l of BMG (100 mM potassium phosphate buffer (pH=6.0), 1.34% yeast nitrogen base,  $4 \times 10^{-5}$  % biotin, 1% glycerol) supplemented with 0.5% cas amino acids and incubated at 1,000 rpm and 30°C for forty-eight hours. Cells were then pelleted by centrifugation at 4,500xg for 5 minutes, and resuspended in 1% BMM (100mM potassium phosphate buffer (pH=6.0), 1.34% yeast nitrogen base,  $4 \times 10^{-5}$  % biotin, 1% methanol)

for induction during seventy-two hours at 1,000 rpm and 20°C. The protein secreted to the extracellular media was then analyzed via SDS-PAGE, ELISA, and Western Blot.

**[0139]** All of the methods disclosed and claimed herein can be made and executed without undue experimentation in light of the present disclosure. While the compositions and methods disclosed herein have been described in terms of certain embodiments, it will be apparent to those of skill in the art that variations may be applied to the methods and in the steps or in the sequence of steps of the method described herein without departing from the concept, spirit and scope of the disclosed embodiments. More specifically, it will be apparent that certain agents which are both chemically and physiologically related may be substituted for the agents described herein while the same or similar results would be achieved. All such similar substitutes and modifications apparent to those skilled in the art are deemed to be within the spirit, scope and concept of the embodiments disclosed herein as defined by the appended claims.

## REFERENCES

**[0140]** The following references, to the extent that they provide exemplary procedural or other details supplementary to those set forth herein, are specifically incorporated herein by reference.

Bernauer et al., *Komagataella phaffii* as emerging model organism in fundamental research. *Front. Microbiol.* (January 11, 2021).

Besada-Lombana & Da Silva (2019) Engineering the early secretory pathway for increased protein secretion in *Saccharomyces cerevisiae*. *Metabolic Engineering*, 55, 142-151 (September 2019).

Dalvie et al. (2020) "Host-informed expression of CRISPR guide RNA for genomic engineering in *Komagataella phaffii*." *ACS Synth. Biol.*, 9(1), 26-35 (December 11, 2019).

Duran & Kahve (2017) The use of lactoferrin in food industry. *Academic Journal of Science*, 07(02), 89-94.

Heiss et al. (2015) Multi-step processing of the secretion leader of the extracellular protein Epx1 in *Pichia pastoris* and implications for protein localization. *Microbiology*, 161(7) (July 1, 2015).

Madden, Tolstorukov, & Cregg, Book Chapter: Electroporation of *Pichia pastoris*. *Genetic Transformation Systems 87 in Fungi*, Volume 1, Fungal Biology. M.A. van den Berg and K. Maruthachalam (eds.) (2014).

Nicholl, *An Introduction to Genetic Engineering*. 2nd edition (Cambridge: Cambridge University Press, 2002), Glossary.

*Recombinant Protein Production in Yeast*, Brigitte Gasser & Diethard Mattanovich (eds.) (Springer, 2019).

U.S. Pat. No. 4,977,137 (Nicols et al.)

U.S. Pat. No. 5,571,691 (Conneely et al.)

U.S. Pat. No. 7,335,512 (Callewaert et al.)

U.S. Pat. No. 7,344,867 (Connolly)

U.S. Pat. No. 7,749,960 (Vidal et al.)

U.S. Pat. No. 7,524,815 (Vidal et al.)

U.S. Pat. No. 7,914,822 (Medo)

U.S. Pat. No. 8,440,456 (Callewaert et al.)

U.S. Pat. No. 8,871,445 (Cong et al.)

U.S. Pat. No. 8,802,650 (Buck et al.)

U.S. Pat. No. 8,821,878 (Medo et al.)

U.S. Pat. No. 8,927,027 (Fournell et al.)

5 U.S. Pat. No. 7,449,308 (Gerngross et al.)

U.S. Pat. Publ. 2012/0142580 (Nutten et al.)

## EMBODIMENTS

10

### [0141]

15

**Embodiment 1:** An isolated nucleic acid encoding a polypeptide comprising a sequence having at least 90% sequence identity to SEQ ID NO:1, 2, 3, or 4.

**Embodiment 2:** The isolated nucleic acid of embodiment 1, wherein the sequence comprises SEQ ID NO:1, 2, 3, or 4.

20

**Embodiment 3:** The isolated nucleic acid of embodiment 1 or 2, wherein the polypeptide further comprises a sequence of a mammalian protein.

**Embodiment 4:** The isolated nucleic acid of embodiment 3, wherein the mammalian protein is a human milk protein.

25

**Embodiment 5:** The isolated nucleic acid of embodiment 4, wherein the human milk protein is secretory IgA (sIgA), xanthine dehydrogenase, lactoferrin, lactoperoxidase, butyrophilin, lactadherin, adiponectin,  $\beta$ -casein,  $\kappa$ -casein, leptin, lysozyme, or  $\alpha$ -lactalbumin.

**Embodiment 6:** The isolated nucleic acid of embodiment 5, wherein the human milk protein is human lactoferrin.

30

**Embodiment 7:** The isolated nucleic acid of any of embodiments 1-6, wherein the sequence has at least 90% sequence identity to SEQ ID NO:1.

**Embodiment 8:** The isolated nucleic acid of any one of embodiments 1-6, wherein the sequence comprises SEQ ID NO:1.

35

**Embodiment 9:** The isolated nucleic acid of embodiment 8, wherein the isolated nucleic acid comprises a nucleic acid sequence having at least 80% identity to SEQ ID NO:41.

**Embodiment 10:** The isolated nucleic acid of embodiment 9, wherein the nucleic acid sequence comprises SEQ ID NO:41.

40

**Embodiment 11:** The isolated nucleic acid of embodiment 8, wherein the polypeptide comprises SEQ ID NO:5.

**Embodiment 12:** The isolated nucleic acid of embodiment 11, wherein the isolated nucleic acid comprises a nucleic acid sequence having at least 80% identity to SEQ ID NO:46.

45

**Embodiment 13:** The isolated nucleic acid of embodiment 12, wherein the nucleic acid sequence comprises SEQ ID NO:46.

**Embodiment 14:** The isolated nucleic acid of any of embodiments 1-6, wherein the sequence has at least 90% sequence identity to SEQ ID NO:2.

50

**Embodiment 15:** The isolated nucleic acid of any one of embodiments 1-6, wherein the sequence comprises SEQ ID NO:2.

55

**Embodiment 16:** The isolated nucleic acid of embodiment 15, wherein the isolated nucleic acid comprises a nucleic acid sequence having at least 80% identity to SEQ ID NO:42.

**Embodiment 17:** The isolated nucleic acid of embodiment 16, wherein the nucleic acid sequence comprises SEQ

ID NO:42.

**Embodiment 18:** The isolated nucleic acid of embodiment 15 wherein the polypeptide comprises SEQ ID NO:6.

**Embodiment 19:** The isolated nucleic acid of embodiment 18, wherein the isolated nucleic acid comprises a nucleic acid sequence having at least 80% identity to SEQ ID NO:47.

**Embodiment 20:** The isolated nucleic acid of embodiment 19, wherein the nucleic acid sequence comprises SEQ ID NO:47.

**Embodiment 21:** The isolated nucleic acid of any of embodiments 1-6, wherein the sequence has at least 90% sequence identity to SEQ ID NO:3.

**Embodiment 22:** The isolated nucleic acid of any one of embodiments 1-6, wherein the sequence comprises SEQ ID NO:3.

**Embodiment 23:** The isolated nucleic acid of embodiment 22, wherein the isolated nucleic acid comprises a nucleic acid sequence having at least 80% identity to SEQ ID NO:43.

**Embodiment 24:** The isolated nucleic acid of embodiment 23, wherein the nucleic acid sequence comprises SEQ ID NO:43.

**Embodiment 25:** The isolated nucleic acid of embodiment 22, wherein the polypeptide comprises SEQ ID NO:7.

**Embodiment 26:** The isolated nucleic acid of embodiment 25, wherein the isolated nucleic acid comprises a nucleic acid sequence having at least 80% identity to SEQ ID NO:48.

**Embodiment 27:** The isolated nucleic acid of embodiment 26, wherein the nucleic acid sequence comprises SEQ ID NO:48.

**Embodiment 28:** The isolated nucleic acid of any of embodiments 1-6, wherein the sequence has at least 90% sequence identity to SEQ ID NO:4.

**Embodiment 29:** The isolated nucleic acid of any one of embodiments 1-6, wherein the sequence comprises SEQ ID NO:4.

**Embodiment 30:** The isolated nucleic acid of embodiment 29, wherein the isolated nucleic acid comprises a nucleic acid sequence having at least 80% identity to SEQ ID NO:44.

**Embodiment 31:** The isolated nucleic acid of embodiment 30, wherein the nucleic acid sequence comprises SEQ ID NO:44.

**Embodiment 32:** The isolated nucleic acid of embodiment 32, wherein the polypeptide comprises SEQ ID NO:8.

**Embodiment 33:** The isolated nucleic acid of embodiment 32, wherein the isolated nucleic acid comprises a nucleic acid sequence having at least 80% identity to SEQ ID NO:49.

**Embodiment 34:** The isolated nucleic acid of embodiment 33, wherein the nucleic acid sequence comprises SEQ ID NO:49.

**Embodiment 35:** A vector comprising the nucleic acid of any of embodiments 1-34.

**Embodiment 36:** An engineered eukaryotic cell comprising the nucleic acid of any of embodiments 1-34 or the vector of embodiment 35.

**Embodiment 37:** The engineered eukaryotic cell of embodiment 36, wherein the cell is a fungal cell.

**Embodiment 38:** The engineered eukaryotic cell of embodiment 37, wherein the fungal cell is a *Arxula*, *Aspegillus*,

*Aurantiochytrium, Candida, Claviceps, Cryptococcus, Cunninghamella, Geotrichum, Hansenula, Kluyveromyces, Kodamaea, Komagataella, Leucosporidiella, Lipomyces, Mortierella, Ogataea, Pichia, Prototheca, Rhizopus, Rhodospiridium, Rhodotorula, Saccharomyces, Schizosaccharomyces, Tremella, Trichosporon, Wickerhamomyces, or Yarrowia* cell.

**Embodiment 39:** The engineered eukaryotic cell of embodiment 38, wherein the cell is a yeast cell.

**Embodiment 40:** The engineered eukaryotic cell of embodiment 39, wherein the yeast cell is a *Komagataella* cell.

**Embodiment 41:** The engineered eukaryotic cell of embodiment 40, wherein the yeast cell is a *Komagataella phaffii*, *Komagataella pastoris*, or *Komagataella pseudopastoris* cell.

**Embodiment 42:** The engineered eukaryotic cell of any one of embodiments 36-41, wherein the nucleic acid is integrated into the genome of the cell.

**Embodiment 43:** The engineered eukaryotic cell of any one of embodiments 36-41, wherein the nucleic acid is not integrated into the genome of the cell.

**Embodiment 44:** A method for producing a secreted protein, the method comprising growing the cell of any one of embodiments 36-43 under conditions sufficient to secrete the polypeptide from the cell.

**Embodiment 45:** The method of embodiment 44, further comprising collecting the secreted protein.

**Embodiment 46:** The method of embodiment 44 or 45, wherein the secreted protein is a human milk protein.

**Embodiment 47:** The method of embodiment 46, wherein the human milk protein is secretory IgA (sIgA), xanthine dehydrogenase, lactoferrin, lactoperoxidase, butyrophilin, lactadherin, adiponectin,  $\beta$ -casein,  $\kappa$ -casein, leptin, lysozyme, or  $\alpha$ -lactalbumin.

**Embodiment 48:** The method of any one of embodiments 44-47, wherein the human milk protein comprises one or more human-like N-glycans.

**Embodiment 49:** The method of any one of embodiments 44-48, further comprising generating a mixture comprising the human milk protein and one or more components of an infant formula.

**Embodiment 50:** An engineered yeast cell comprising a nucleic acid encoding a polypeptide comprising a sequence having at least 90% sequence identity to SEQ ID NO:1, 2, 3, or 4.

**Embodiment 51:** The engineered yeast cell of embodiment 50, wherein the sequence comprises SEQ ID NO:1, 2, 3, or 4.

**Embodiment 52:** The engineered yeast cell of embodiment 51, wherein the sequence comprises SEQ ID NO:3.

**Embodiment 53:** The engineered yeast cell of any one of embodiments 50-52, wherein the polypeptide further comprises a sequence of a mammalian protein.

**Embodiment 54:** The engineered yeast cell of embodiment 53, wherein the mammalian protein is a human milk protein.

**Embodiment 55:** The engineered yeast cell of embodiment 54, wherein the human milk protein is secretory IgA (sIgA), xanthine dehydrogenase, lactoferrin, lactoperoxidase, butyrophilin, lactadherin, adiponectin,  $\beta$ -casein,  $\kappa$ -casein, leptin, lysozyme, or  $\alpha$ -lactalbumin.

**Embodiment 56:** The engineered yeast cell of embodiment 55, wherein the human milk protein is human lactoferrin.

**Embodiment 57:** An engineered yeast cell comprising:

(a) a first nucleic acid encoding a polypeptide comprising:

- (i) a sequence having at least 90% sequence identity to SEQ ID NO:1, 2, 3, or 4 and
- (ii) a sequence of a human milk protein; and

(b) a second nucleic acid encoding an  $\alpha$ -1,2-mannosidase (Man-I) protein,

wherein the cell does not express a functional OCH1 protein.

**Embodiment 58:** The engineered yeast cell of embodiment 57, wherein the sequence of (i) comprises SEQ ID NO:1, 2, 3, or 4.

**Embodiment 59:** The engineered yeast cell of embodiment 57 or 58, wherein the human milk protein is secretory IgA (sIgA), xanthine dehydrogenase, lactoferrin, lactoperoxidase, butyrophilin, lactadherin, adiponectin,  $\beta$ -casein,  $\kappa$ -casein, leptin, lysozyme, or  $\alpha$ -lactalbumin.

**Embodiment 60:** The engineered yeast cell of embodiment 57, wherein the human milk protein is human lactoferrin.

**Embodiment 61:** The engineered yeast cell of embodiment 57, wherein the human milk protein is human  $\alpha$ -lactalbumin.

**Embodiment 62:** The engineered yeast cell of any one of embodiments 57-61, wherein the Man-I protein is fused to a HDEL C-terminal tag.

**Embodiment 63:** The engineered yeast cell of any one of embodiments 57-62, further comprising a third nucleic acid encoding one or more of:

- (a) a N-acetylglucosaminyltransferase I (GnT-I) protein;
- (b) an  $\alpha$ -1,3/6-Mannosidase (Man-II) protein;
- (c) a  $\beta$ -1,2-acetylglucosaminyltransferase (GnT-II) protein; and
- (d) a  $\beta$ -1,4-galactosyltransferase (GalT) protein.

**Embodiment 64:** An infant formula comprising a human glycoprotein having human-like N-linked glycosylation.

**Embodiment 65:** The infant formula of embodiment 64, wherein the human glycoprotein is a human milk protein.

**Embodiment 66:** The infant formula of embodiment 65, wherein the human milk protein is secretory IgA (sIgA), xanthine dehydrogenase, lactoferrin, lactoperoxidase, butyrophilin, lactadherin, adiponectin,  $\beta$ -casein,  $\kappa$ -casein, leptin, lysozyme, or  $\alpha$ -lactalbumin.

**Embodiment 67:** The infant formula of embodiment 66, wherein the human milk protein is human lactoferrin.

**Embodiment 68:** The infant formula of any one of embodiments 64-67, wherein the lactoferrin has a glycan pattern that is different from a glycan pattern of any human lactoferrin naturally occurring in human breast milk.

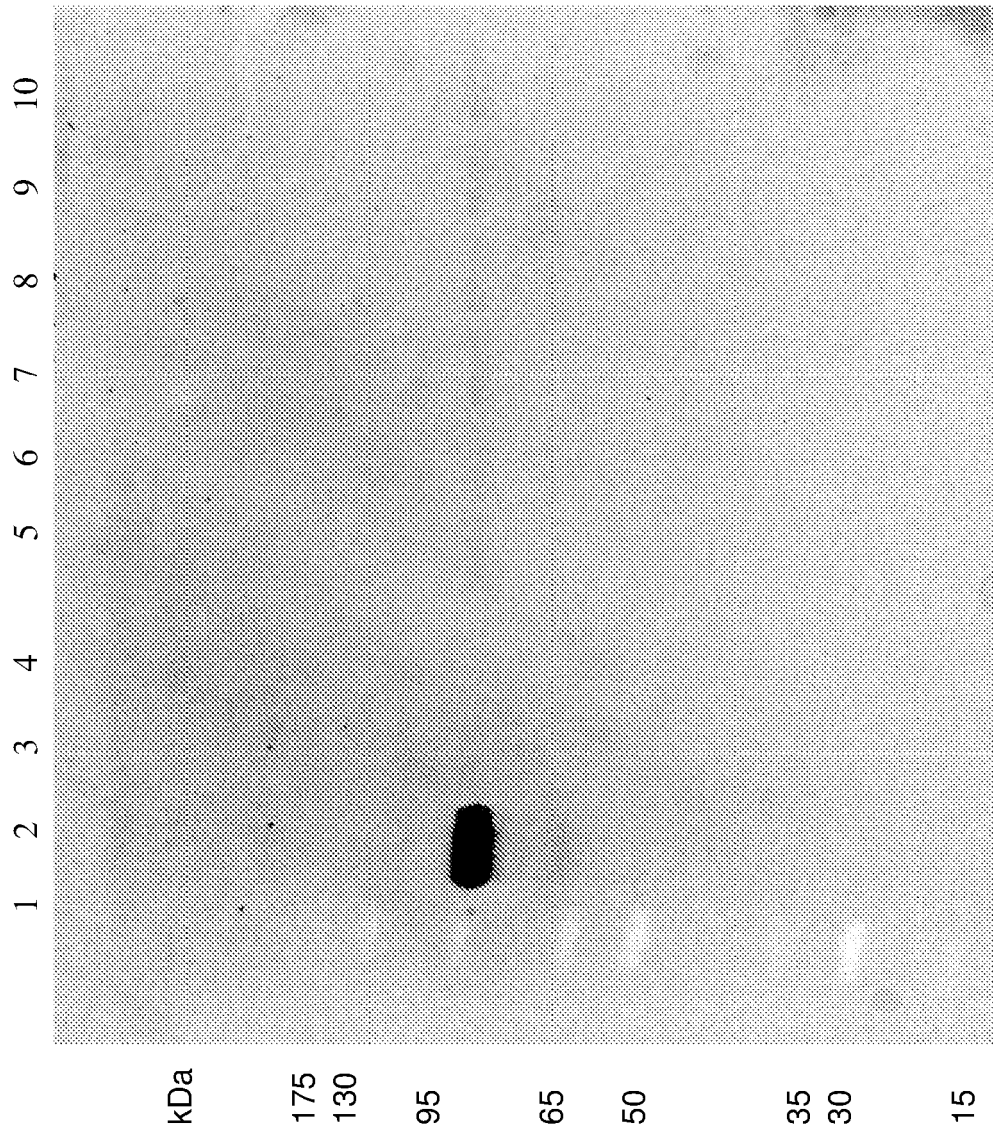
**Embodiment 69:** The infant formula of any one of embodiments 64-68, wherein the human glycoprotein is produced by the method of any one of embodiments 44-49.

## Claims

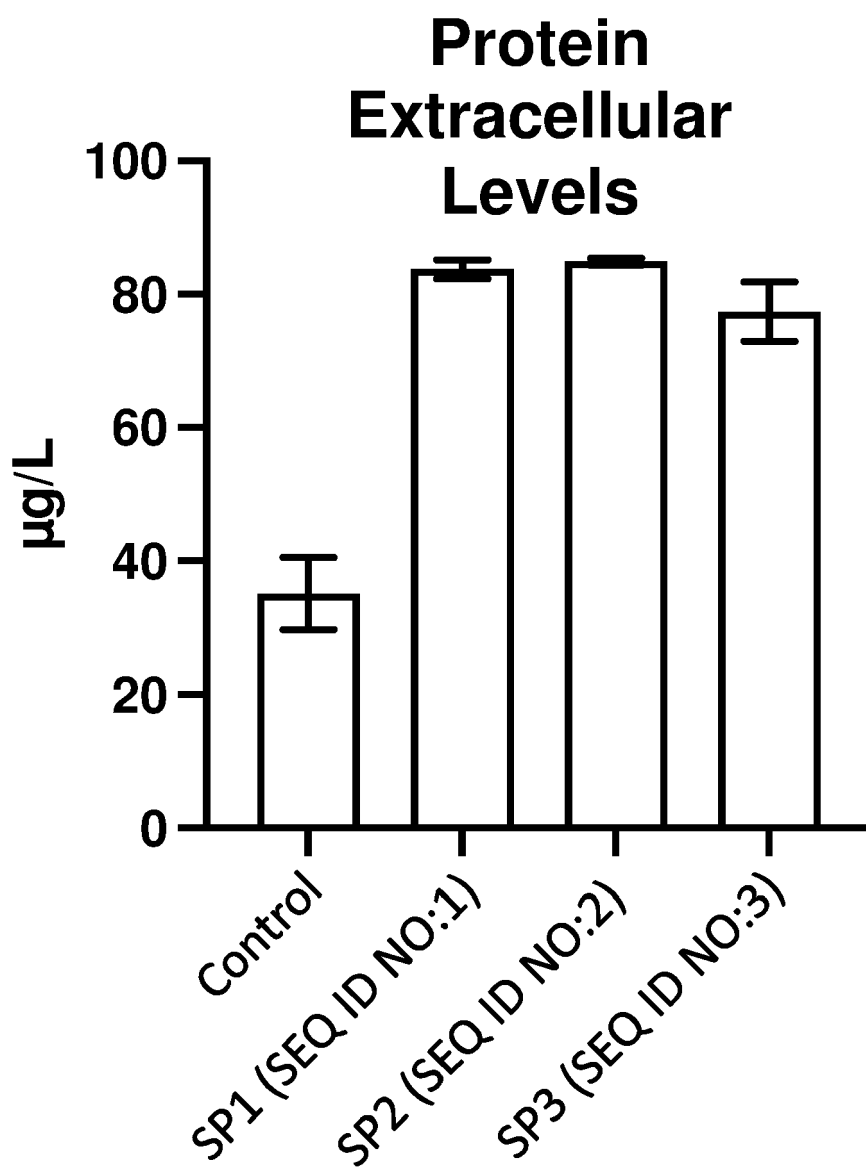
1. An infant formula comprising a recombinant human glycoprotein comprising a glycan of a human glycoprotein, preferably the recombinant human glycoprotein is a human milk protein.
2. The infant formula of claim 1, wherein the glycan is a human-like glycan.
3. The infant formula of any one of claims 1 to 2, wherein the recombinant human glycoprotein is secretory IgA (sIgA), xanthine dehydrogenase, lactoferrin, lactoperoxidase, butyrophilin, lactadherin, adiponectin,  $\beta$ -casein,  $\kappa$ -casein, leptin, lysozyme, or  $\alpha$ -lactalbumin.



4. The infant formula of any one of claims 1 to 3, wherein the recombinant human glycoprotein is human lactoferrin.
5. The infant formula of claim 1, wherein the recombinant human glycoprotein has a glycan pattern different from a glycan pattern of a human lactoferrin naturally occurring in human breast milk.
6. The infant formula of claim 1, wherein the recombinant human glycoprotein is produced by a fungus cell.
7. The infant formula of claim 1, wherein said recombinant human glycoprotein is produced by an *Arxula*, *Aspegillus*, *Aurantiochytrium*, *Candida*, *Claviceps*, *Cryptococcus*, *Cunninghamella*, *Geotrichum*, *Hansenula*, *Kluyveromyces*, *Kodamaea*, *Komagataella*, *Leucosporidiella*, *Lipomyces*, *Mortierella*, *Ogataea*, *Pichia*, *Prototheca*, *Rhizopus*, *Rhodospiridium*, *Rhodotorula*, *Saccharomyces*, *Schizosaccharomyces*, *Tremella*, *Trichosporon*, *Wickerhamomyces* or *Yarrowia* cell.
8. The infant formula of claim 1, wherein said recombinant human glycoprotein is produced by a yeast cell.
9. The infant formula of claim 8, wherein said yeast cell is a *Komagataella* cell, preferably said yeast cell is a *Komagataella phaffii*, *Komagataella pastoris* or *Komagataella pseudopastoris* cell.
10. The infant formula of claim 8, wherein said yeast cell comprises *Pichia pastoris*, preferably *Pichia pastoris* is *Pichia pastoris*, *Komagataella pastoris* or *Komagataella phaffii*.
11. The infant formula of claim 6, wherein said fungus cell comprises an exogenous nucleic acid encoding a polypeptide of a secreted recombinant human glycoprotein, preferably said fungus cell further comprises an exogenous nucleic acid encoding an alpha-1,2-mannosidase (Man-I) protein and/or does not express a functional OCHI protein.
12. The infant formula of claim 11, wherein said exogenous nucleic acid encodes a human glycoprotein comprising a sequence having at least 90% sequence identity to SEQ ID NO:1, 2, 3 or 4.
13. The infant formula of claim 11, wherein said Man-I protein is fused to an HDEL C-terminal tag.
14. The infant formula of claim 11, wherein said fungus cell further comprises an exogenous nucleic acid encoding one or more of:
  - (a) an N-acetylglucosaminyltransferase I (GnT-I) protein;
  - (b) an  $\alpha$ -1,3/6-Mannosidase (Man-II) protein;
  - (c) a  $\beta$ -1,2-acetylglucosaminyltransferase (GnT-II) protein; or
  - (d) a  $\beta$ -1,4-galactosyltransferase (GalT) protein.
15. The infant formula of claim 11, wherein said exogenous nucleic acid is integrated into a genome of said fungus cell or wherein said exogenous nucleic acid is not integrated into a genome of said fungus cell.



**FIG. 1**

**FIG. 2**

## REFERENCES CITED IN THE DESCRIPTION

*This list of references cited by the applicant is for the reader's convenience only. It does not form part of the European patent document. Even though great care has been taken in compiling the references, errors or omissions cannot be excluded and the EPO disclaims all liability in this regard.*

## Patent documents cited in the description

- US 63227820 [0001]
- US 63273858 [0001]
- US 4683202 A [0088]
- US 9617550 B [0109]
- US 20110059502 A [0123]
- US 4977137 A, Nicols [0140]
- US 5571691 A, Conneely [0140]
- US 7335512 B, Callewaert [0140]
- US 7344867 B, Connolly [0140]
- US 7749960 B, Vidal [0140]
- US 7524815 B, Vidal [0140]
- US 7914822 B, Medo [0140]
- US 8440456 B, Callewaert [0140]
- US 8871445 B, Cong [0140]
- US 8802650 B, Buck [0140]
- US 8821878 B, Medo [0140]
- US 8927027 B, Fournell [0140]
- US 7449308 B, Gerngross [0140]
- US 20120142580 A, Nutton [0140]

## Non-patent literature cited in the description

- **BAIROCH A.** The ENZYME database in 2000. *Nucleic Acids Res.*, 01 January 2000, vol. 28 (1), 304-5 [0052]
- **GREEN ; SAMBROOK.** Molecular Cloning: A Laboratory Manual. 2012 [0088]
- *Gene*, 1995, vol. 164, 49-53 [0088]
- **CHEN ; OROZCO.** *Nucleic Acids Research*, 1988, vol. 16, 8411 [0094]
- *Protist*, 2004, vol. 155, 381-93 [0104]
- **MADDEN ; TOLSTORUKOV ; CREGG.** Fungal Biology. *Fungi*, 2014, vol. 1 [0129] [0137]
- **HEISS et al.** *Microbiology*, 2015, vol. 161 (7 [0130]
- **BERNAUER et al.** Komagataella phaffii as emerging model organism in fundamental research. *Front. Microbiol.*, 11 January 2021 [0140]
- **BESADA-LOMBANA ; DA SILVA.** Engineering the early secretory pathway for increased protein secretion in *Saccharomyces cerevisiae*. *Metabolic Engineering*, September 2019, vol. 55, 142-151 [0140]
- **DALVIE et al.** Host-informed expression of CRISPR guide RNA for genomic engineering in *Komagataella phaffii*. *ACS Synth. Biol.*, 11 December 2019, vol. 9 (1), 26-35 [0140]
- **DURAN ; KAHVE.** The use of lactoferrin in food industry. *Academic Journal of Science*, 2017, vol. 07 (02), 89-94 [0140]
- **HEISS et al.** Multi-step processing of the secretion leader of the extracellular protein Epx1 in *Pichia pastoris* and implications for protein localization. *Microbiology*, 01 July 2015, vol. 161 (7 [0140]
- Electroporation of *Pichia pastoris*. Genetic Transformation Systems 87. **MADDEN ; TOLSTORUKOV ; CREGG.** *Fungi*. 2014, vol. 1 [0140]
- **NICHOLL.** An Introduction to Genetic Engineering. Cambridge University Press, 2002 [0140]
- Recombinant Protein Production in Yeast. Springer, 2019 [0140]